Molecular evolution in two species of *Drosophila*

Sophie Marion de Procé

PhD

The University of Edinburgh

2010

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text.

This work has not been submitted for any other degree of professional qualification except as specified.

Sophie Marion de Procé, March 19, 2010.

Acknowledgements

This work would not have been possible without the help and the encouragement of a large number of people. I am very grateful to all of the following:
Brian Charlesworth for all the supervision, encouragement and useful comments on the manuscript and thesis chapters, and for his patience. Peter Keightley for additional supervision and comments on the manuscript.
Andrea Betancourt for my training in the molecular and fly lab work and encouragement throughout, for invaluable help in the laboratory, as well as with fly crosses and in answering various questions more or less theory-related and comments on some chapters; Daniel Halligan for his patience and great help with Perl programs, Kai Zeng for his help in statistical analyses, Penny Haddrill for advice on splicing sites removal and intron alignment and for letting me use her *D. simulans* data, and Xulio Maside for his primers.
Thanks to my friends in Edinburgh, especially Toni, Jim, Cecile, Kaidy, Farina, Lindsay and most of all Iain for his support in the last months.
Finally thanks to my parents for bringing me up to be interested in biology and believing in me, and my siblings Thibault and Solène for all their support and encouragement.

This work was conducted as part of the GENACT Project, funded by the Marie Curie Host Fellowships for Early Stage training to SMP, as part of the 6th Framework Programme of the European Commission. BC was supported by the Royal Society. BAC sequencing and DLH were funded by a Wellcome Trust and BBSRC grant to PDK and BC. We thank Dr. Mark Dorris for isolating BACs and Dr. Jane Rogers at the Wellcome Trust Sanger Institute for organizing their sequencing.

Publications

The following published paper has arisen from this thesis.

- Marion de Procé S, Halligan DL, Keightley PD and B Charlesworth. (2009) Patterns of DNA-sequence divergence between *Drosophila miranda* and *D. pseudoobscura*. *J. Mol. Evol.* **69**:601–611.

Abstract

Studying evolution at the level of DNA sequences allows the detection of past and recent natural selection. Natural selection has generally been seen as a force acting on protein-coding nucleotide sequences only. However, a number of studies have recently shown that introns and intergenic sequences can also be subject to natural selection. The main aim of this thesis was to detect natural selection in non-coding sequences using *Drosophila* species, a widely used population genetics model. I have used different methods to determine if evidence for natural selection could be found in two lesser-known species, *Drosophila americana* and *Drosophila miranda*.

In Chapter 2, I obtained sequences for a large number of genes in *D. miranda* from BAC sequences, and compared these with sequences from its close relative, *D. pseudoobscura*. As in previous studies in *D. melanogaster*, I found a negative relationship between intron length and intron divergence, suggesting that longer introns are under selective constraint. I also found a negative correlation between the rate of non-synonymous substitutions and codon usage bias, suggesting that fast-evolving genes have a lower codon usage bias, consistent with strong positive selection interfering with weak selection for codon usage.

Secondly, in Chapter 3, I gathered polymorphism data for a smaller number of genes in *D. americana* in order to distinguish between positive and negative selection using methods that require polymorphism and divergence. I found that introns are subject to similar evolutionary forces as synonymous sites. I failed to detect a significant relationship between intron length and divergence or polymorphism. Surprisingly, the direction of this relationship seems to be the opposite of that in previous findings, with longer introns being more diverged than smaller introns. First introns show lower polymorphism and divergence than non-first introns, suggesting that they may be more constrained, although the difference is not significant.

Using the same *D. americana* dataset, I then focussed, in Chapter 4, on insertions and deletions to test the hypothesis that insertions are favoured to compensate for the

deletion bias in *Drosophila*. I used a maximum-likelihood method that takes into account demographic history, in this case a recent population expansion and then calculates the selection coefficients. Although it was not significant, the values suggest positive selection acting on insertions, as expected.

In Chapter 5, using the same maximum-likelihood method, I looked at GC to AT polymorphisms in the *D. americana* intron dataset. It is expected to observe as many GC to AT changes as AT to GC changes and similar mean frequencies if no selection is acting. I find evidence for a preference for GC in introns in my dataset. I also investigated codon usage bias using preferred and unpreferred codons changes and results suggest that there is selection for codon usage bias. Using LDhat on the *D. americana* dataset, I find that recombination estimates are not significantly different between introns and coding sequences, which is of significance in relation to interpretations of differences in the apparent strength of selection on non-coding and synonymous sites.

Finally, in Chapter 6, I looked at a factor that can affect natural selection: gene expression. I used gene expression data from seven *Drosophila* species to test the hypothesis that genes on the 4^{th} chromosome or Muller element F, which has low crossing-over, have higher gene expression than genes on other chromosomes as previously found. I find that microarray data yields opposite results to the EST data, suggesting that gene expression is actually lower on Muller element F.

Contents

Abbreviations

List of commonly used abbreviations:

bp	Base Pairs
kb	Kilobase
µl	Microlitre
ml	Millilitres
mRNA	Messenger RiboNucleic Acid
PCR	Polymerase Chain Reaction
SE	Standard Error
Taq	*Thermus aquaricus* polymerase
UTR	Untranslated Transcribed Regions

All other abbreviations are for chemical formulae or are detailed in the main text.

1 Introduction

1.1 Evolution of non-coding DNA

A large proportion of the euchromatic portion of the *Drosophila* genome is comprised of non-coding sequence that is not translated into proteins. For example, 80% of the 120 Mb euchromatic part of the *Drosophila melanogaster* genome is non-coding (Adams et al. 2000). Non-coding sequences are generally divided into two categories: introns within genes, and intergenic sequences between genes. Intergenic sequences can contain transposable elements, regulatory sites and pseudogenes created by gene duplication, as well as noncoding RNA and matrix scaffold attachment sequences. Untranslated regions (UTR) flank coding sequences at the 5' or 3' ends and belong to the gene mRNA sequence, therefore they need to be annotated in genomes to distinguish them from intergenic sequence. For a long time, non-coding sequence has been thought to lack function and to accumulate mutations faster than synonymous sites in coding sequences, being only subject to genetic drift and mutation, and possibly mutational repair processes such as biased gene conversion. The term "junk" DNA has even been widely used to describe this type of sequence.

However, recent evolutionary studies have shown, using different approaches, that a high fraction of non-coding DNA is selectively constrained in *Drosophila* species (see §1.3 below), suggesting that a large fraction of these non-coding sequences have some function. As we have begun to see that non-coding sequences are under extensive selective constraints (Bergman and Kreitman 2001), questions arise as to what functions they might have. Roles in pre-mRNA secondary structure (Kirby et al. 1995; Leicht et al. 1995; Chen and Stephan 2003, Rogic et al. 2008), gene regulation (Arnone and Davidson 1997; Hardison 2000; Parsch 2004) or RNA editing (Reenan 2005) have been suggested, and there is increasing experimental evidence to support these (Birney et al. 2007). Ultraconserved sequences between humans and rodents also indicate a role in transcriptional regulation and development, specifically in the development of the

nervous system (Visel et al. 2008). More recently, it has also been shown that natural selection on coding and non-coding DNA sequences is related to nucleosome organization (Warnecke et al. 2008; Babbitt and Kim 2008; Kaplan et al. 2009). The functionality of non-coding sequences implies that they can be subject to selective forces such as purifying selection removing deleterious mutations, or positive selection fixing advantageous mutations.

1.2 Detecting natural selection in DNA sequences

Natural selection eliminates deleterious mutations, which results in the conservation of sequences between different species and a skew towards rare alleles (Kimura 1968), or favours advantageous mutations, which results in higher divergence between species (Gillespie 1991; Kimura 1983; reviewed in Nielsen 2005). Various methods can be used to detect the action of natural selection at the sequence level, using different parts of the data.

1.2.1 Comparative genomics

One way of determining the amount of non-coding DNA subject to natural selection is through comparative genomics. The availability of several complete genomes makes this method a thorough and promising approach. Under the neutral theory of molecular evolution, it is predicted that functionally important parts of the genome will evolve more slowly than those lacking function (Kimura 1983). Therefore, by comparing genomes of different species, we can find regions of very high similarity, indicative of selective constraint (Bergman and Kreitman 2001; Stark et al. 2007; Stone et al. 2005). This method gives an overall picture and can therefore point directly to low-divergence regions of the genome. Simply comparing genomes, however, can be biased by several factors. First, apparent sequence conservation can be due to lower mutation rates, potentially caused by differences in base composition, so that any comparison needs to account for this. It has also been shown that regulatory sequences do not seem well

conserved as a class over long periods of evolution (Richards et al. 2005), suggesting that the function of cis-regulatory elements can be conserved without the primary sequence being conserved (Ludwig et al. 2000). This means that functionality can be underestimated by looking at conserved sequences only. However, Halligan et al. (2006) have shown that non-coding sequences are highly conserved between two closely related species, *D. melanogaster* and *D. simulans*, with levels of constraint similar to those for amino-acid sites.

Functional sequences can also be subject to positive selection, which increases the observed level of divergence. A widely used criterion for detecting natural selection in coding sequences using divergence is the ratio of non-synonymous to synonymous substitutions. A ratio of 1 is expected under neutral evolution with no selective constraint, an excess of non-synonymous substitutions indicates positive selection, whereas an excess of synonymous substitutions shows purifying selection. In the case of non-coding DNA, divergence for putatively constrained sequence can be compared with divergence for an unconstrained standard, such as that for fourfold degenerate synonymous sites, and selective constraint can be measured in this way ($C = 1 - O/E$, where E is the expected divergence, and O is the observed divergence; Halligan et al. 2004).

1.2.2 Methods using polymorphism data

The main advantage of using polymorphism data is to discriminate between lower constraint and positive selection. It should be noted that there are ways of doing this using phylogenetic methods (PAML: Yang 1997; Yang 2007). Polymorphism data also allows discrimination between fixed differences and polymorphic sites among the observed substitutions. Indeed, published genome sequences only come from a single individual, which may not be representative of the whole species, and can lead to biased estimates of divergence. Using polymorphism data also allows using more elaborate tests to detect natural selection, as discussed below, but it often reduces the amount of sequence that can be studied. There is thus a trade-off between the extent of the sequence available for analysis and the additional information provided by polymorphism data in analyses. However, new methods are being developed that will allow fast re-sequencing

of whole genomes (e.g. 454 Life Sciences, Solexa), which should make it possible to have polymorphism data on a much larger scale and at lower costs (Hall 2007).

Some tests commonly used to detect natural selection from polymorphism data are Tajima's *D* (Tajima 1989), Fu and Li's *D* (Fu and Li 1993), the Hudson-Kreitman-Aguadé (HKA; Hudson et al. 1987) and the McDonald-Kreitman (MK; McDonald and Kreitman 1991) tests. Tajima's *D* (Tajima 1989) compares two estimates of DNA sequence variation: θ, using *S*, the number of segregating sites, and π, a measure of nucleotide diversity that uses the average pairwise differences. These estimates should be similar under neutrality. Under constant demographic parameters, an excess of intermediate frequency alleles ($D > 0$) might be due to balancing selection (eg. heterozygote advantage), whereas an excess of rare alleles ($D < 0$) can indicate either purifying selection or a nearby selective sweep. This approach, however, assumes that the population has been demographically stable for a long time, because a positive Tajima's *D* can also reflect population bottlenecks (Maruyama and Fuerst, 1985), and a negative Tajima's *D* can be caused by population expansion (Maruyama and Fuerst, 1984).

Fu and Li's *D* (1993) is similar to Tajima's *D* but is based on the assumption that the expected number of derived mutations that are present only once in a sample, η_e, is equal to θ. Fu and Li's *D* shares much information with Tajima's *D*, and may only be more sensitive than Tajima's *D* in some population genetic scenarios such as selective sweeps, that tend to generate an excess of singletons.

Both the HKA (Hudson et al. 1987) and the McDonald-Kreitman (McDonald and Kreitman, 1991) tests use the expectations that the measure of genetic diversity θ equals $4N_e\mu$ under neutrality, all loci share the same effective population size N_e, and each locus has its own characteristic neutral mutation rate μ.

Under the neutral theory, polymorphism and divergence are expected to be correlated, because they share a neutral mutation rate. The HKA test requires intraspecific polymorphism from two or more loci to estimate θ, and interspecific divergence to estimate μ. Selection is inferred when μ and θ vary in non-corresponding fashions between loci, meaning that the relative amounts of polymorphism and divergence vary between loci; it is generally most useful to compare a potentially

selected locus to neutrally evolving loci. If most tests of the potentially selected locus against neutrally evolving loci are significant, it is then likely that this locus has been subject to natural selection.

The McDonald-Kreitman test compares the ratio of polymorphism to divergence for two classes of sites, one of which is potentially selected while the other class is putatively neutrally evolving. These two ratios should be equal under neutrality, so selection is inferred when these ratios differ. If the ratio of potentially selected to putatively neutrally evolving sites is higher between species (divergence) than within species (polymorphism), it is inferred that positive selection has promoted the fixation of potentially selected changes (Eyre-Walker 2006). The converse would mean that there is an excess of polymorphism in the potentially selected class, suggesting the action of weak purifying selection or balancing selection, or a relaxation of selection.

1.3 Non-coding DNA findings in *Drosophila*

1.3.1 Selection on introns

Haddrill et al. (2005) inferred from the patterns of divergence within long introns that constrained sequences were not clustered but uniformly spread across long introns. These sequences would thus be more likely to be regulatory elements involved in precursor messenger RNA (pre-mRNA) secondary structure. However Bergman and Kreitman (2001) found that constrained sequence occurs in blocks that are located throughout the non-coding sequence, with similar levels of constraint in introns and intergenic sequences at short as well as long distance from coding sequence. Halligan and Keightley (2006) found that substitutions between *D. melanogaster* and *D. simulans* are clustered, which means that there are blocks of constrained sequences, confirming results for *Drosophila* of Dermitzakis et al. (2003) and Bergman et al. (2002). This suggests that they control gene expression. However, it is probable that constrained non-coding DNA includes several different types of distribution corresponding to different functions. Bergman and Kreitman (2001) also suggested that constraint on both point substitutions and indels

would indicate the presence of transcription factor binding sites on conserved non-coding sequences.

Haddrill et al. (2005) and Halligan and Keightley (2006) found that the negative correlation between divergence and intron length is significant for both first and non-first introns. A similar result was also found in a study by Bachtrog and Andolfatto (2006). Bergman and Kreitman (2001) found that long introns (> 80 bp) have constrained blocks and that divergence still decreases after this limit. Longer introns, especially first introns, may therefore contain more regulatory elements. Using 225 intron fragments on X and autosomes, with various lengths and positions in genes, Haddrill et al. (2005) measured divergence between introns of two closely related *Drosophila* species, *D. melanogaster* and *D. simulans*, taking into account GC content and intron length. Haddrill et al. (2005) and Marais et al. (2005) showed that longer introns are more conserved. First introns are generally longer than non-first introns in mammals (Smith 1988) and in *Drosophila* (Maroni 1994). They are also more conserved in mammals, as they contain more regulatory elements (Majewski and Ott 2002; Keightley and Gaffney 2003; Chamary and Hurst 2004), and Duret (2001) and Marais et al. (2005) suggest that it could also be the case in *Drosophila*.

Other variables seem to correlate with sequence length that could interfere with divergence. For example, Comeron and Kreitman (2000) found a negative correlation between intron length and local recombination rate. They suggest that this might be due to either selection to maintain the minimal intron size or to selection favouring increased recombination between adjacent exons in low recombination regions (Hill-Robertson effect; Hill and Robertson 1966). Carvalho and Clark (1999) also found an overall negative correlation between intron length and recombination rate, and a positive correlation for short introns. They interpret this differently from Comeron and Kreitman (2000). They suggest that intermediate intron lengths are favoured and so very long and very short introns are deleterious and only occur in low recombination regions, where natural selection is less efficient. Marais et al. (2005) found a negative correlation between intron length and gene expression level, but a positive correlation when only first introns are considered. This seems to confirm the presence of more regulatory

elements controlling gene expression in first introns. Another possible explanation for selection on introns is the importance of their structure. For example, Chen and Stephan (2003) found that mutations in the first intron of *Drosophila Adh* disrupt its structure and might be causing a reduction in splicing efficiency.

Haddrill et al. (2005) found a negative correlation between intron divergence and GC content and Casillas et al. (2007) also observed a higher GC content in conserved non-coding sequences (CNSs). This could be due either to local variation in mutational bias, or selection or biased gene conversion favouring GC over AT, as has been found in a study on *D. simulans* (Haddrill and Charlesworth 2008). Casillas et al. (2007) argue that GC nucleotides are preserved by purifying selection in CNSs. Using predictions from the standard model of drift and reversible mutation, Haddrill et al. (2005) show that local variation in mutational bias might be sufficient to explain this correlation.

1.3.2 Selection for compact genome size?

The observation of high deletion rates for inactive transposable elements in *Drosophila* (Petrov et al. 1996; Petrov and Hartl 1998) and the fact that its genome size is relatively small has led to the hypothesis that there may be selection for a compact genome size (Charlesworth, 1996). The observation of a high proportion of constrained non-coding sequence in *Drosophila* supports this hypothesis (Casillas et al. 2007); this is indeed what we would expect after a period of selection for compact genome size eliminating non-functional sequence. However, Petrov (2002) and Gregory (2004) argue that selection for compact genome size is unlikely to be strong enough to fully explain the observed deletion bias. A way to test this would be to estimate the strength of selection necessary to eliminate the majority of sequence contained in inactive transposable elements and compare it with observed selection coefficients. Another explanation for the removal of inactive transposable elements would be that they interrupt functional non-coding sequence and are therefore selected against because of this, instead of just a lack of function.

In the papers on non-coding DNA that have been discussed so far, constraint was mainly studied through point substitutions and sequence length as opposed to insertions

and deletions (indels). Other work has investigated patterns of selection of indels in introns. In a study on *Arabidopsis lyrata* and *A. thaliana*, Wright et al. (2002) found that introns are consistently shorter in *A. thaliana* due to the accumulation of small indels and potentially reflecting natural selection maintaining large introns in *A. lyrata*. A mutational deletion bias has been observed for paralogous divergence in *Drosophila* retrotransposons (Petrov et al., 1996; Petrov and Hartl, 1998), but this could include fixed differences. A polymorphic deletion bias was reported by Comeron and Kreitman (2000) in a study of 31 genomic regions in *Drosophila melanogaster*. Natural selection can either reinforce this bias or oppose it (Comeron and Kreitman, 2000). For example, Presgraves (2006) has shown that X-linked insertions are favoured in *D. melanogaster* but not in *D. simulans*. He suggested that this result might be due to biased gene conversion-gap repair, which is discussed below, or to differences in rates of crossing-over between X and autosomes and between the two species. Bergman et al. (2002) observed that spacing between conserved non-coding sequences (CNCS) is conserved, further suggesting that these sequences as well as the spacer interval sequences between them may experience selective constraint. Using 15 introns in a multi-species study of the *D. melanogaster* subgroup, Parsch (2003) also showed that length constraints seem to differ between introns within the same gene, with short deletions fixed by genetic drift and long insertions fixed by positive selection to restore intron length. Ometto et al. (2005) found a similar polymorphic deletion bias in intronic and intergenic sequences, supporting the hypothesis that insertions are selected to compensate for the deletion bias. This suggests that there could be selection on spacer length in these regions or on minimal intron size in introns; therefore, it is interesting to examine the nature of selection on indels.

1.3.3 Patterns of evolution in other non-coding DNA

Bergman and Kreitman (2001) studied 40 loci spanning over 100 kb, likely to contain cis-regulatory elements in *D. virilis* and *D. melanogaster*, two relatively distant species. They looked for conserved blocks and measured block length, number of substitutions and insertions-deletions (indel) length. On an even larger scale, Halligan and Keightley (2006) used a whole-genome approach to compare the genome sequences of *D. melanogaster* and *D. simulans*. They compared substitution rates in non-coding sequences with those in putatively unconstrained sites: fast-evolving intron sites and fast-evolving fourfold synonymous sites. They also measured the distance from coding sequence, and intron and intergenic sequence length.

Bergman and Kreitman (2001) found high levels of constraint on both indels and point substitutions, with similar patterns for intronic and intergenic sequences. Andolfatto (2005) used published polymorphism sequences for 51 intergenic fragments on the X chromosome of *D. melanogaster*. He compared these fragments with fourfold degenerate synonymous sites to estimate selective constraint and found that most intronic sequences, 50% of non-UTR (untranslated transcribed regions) intergenic sites and as many as 60% of sites within UTRs are selectively constrained, even in intergenic sequences that were distant from genes. Halligan and Keightley (2006) found high mean constraint within all categories of non-coding DNA studied, suggesting that over 50% of newly arising mutations are removed by selection in non-coding DNA. Using McDonald-Kreitman tests and derived allele frequency spectra, Casillas et al. (2007) also found that weak purifying selection maintains highly conserved non-coding sequences. The existence of constraint on intron content is additional to the probable selection on minimal intron size for correct splicing (Mount et al., 1992). The similarity of results for intronic and intergenic regions suggests that they are equally functional and subject to similar evolutionary processes. Using HKA and McDonald-Kreitman tests, Andolfatto (2005) found that 20% of intronic and intergenic sequence and 60% of UTRs are under positive selection in his study of *D. melanogaster*. Begun et al. (2007) and Haddrill et al. (2008a) also found evidence for purifying and positive selection in different classes of non-coding DNA in *D. simulans*. Haddrill et al. (2008a) found reduced levels of polymorphism and

divergence indicating purifying selection, as well as a skew towards rare variants resulting in strongly negative values for Tajima's *D*. They also estimated that 45% of divergence in introns and 50-70% in UTRs are fixed by positive selection. Thus there is evidence for positive as well as negative selection, so they should both be considered when studying the functionality of non-coding sequences.

Nelson et al. (2004) used surrogate measurements to identify the regulatory complexity of genes. For example, there are generally more regulatory elements in genes that are expressed in a greater number of tissues. Nelson et al. (2004) found that genes with complex regulation are flanked by significantly more non-coding DNA than genes with simple or housekeeping functions, suggesting the presence of regulatory elements in non-coding DNA. Halligan and Keightley (2006) also found a correlation between divergence and sequence length for intergenic sequence, with a peak of divergence at 500 bp for both 3' and 5' intergenic sequences, suggesting that, consistent with Bergman and Kreitman (2001), similar processes occur in long introns and intergenic sequences over 500 bp.

The positive correlation between divergence and intergenic sequence length up to 500 bp might be explained by the presence of UTRs in intergenic DNA, because UTRs constitute a higher proportion of intergenic sequences in short sequences. Furthermore, when excluding UTRs from the analysis, Halligan and Keightley (2006) found a strong negative correlation between sequence length and divergence.

When conserved sequences are found in GC-rich regions where the mutation rate is not particularly GC-biased, there can be another explanation for the high similarity between sequences. It has been argued that GC content is positively correlated with the level of biased gene conversion (BGC), which in effect conserves DNA sequences in the same way as natural selection would. BGC is gene conversion that is biased in favour of one allele over another. BGC towards GC versus AT occurs in many organisms and might have arisen to compensate the GC to AT mutation bias (Marais 2003; Lynch 2007; Lynch 2010). During recombination, the repair system favours GC base pairs, which increases the GC content of the sequence. This mechanism should have the largest effect on non-coding sites, where selection on codon usage bias cannot interfere. In *D.*

melanogaster, Galtier et al. (2006) found that the mean frequency of AT to GC polymorphisms in non-coding DNA is higher than that of GC to AT polymorphisms, suggesting either a GC-biased allele transmission due to BGC, or a nonstationary evolution of base composition. This result was also confirmed in *D. simulans* (Haddrill et al. 2008a) and can be explained by weak selection to maintain CNSs (Casillas et al. 2007). In *Drosophila*, a positive correlation between recombination and non-coding GC content has been found by Marais et al. (2003), which is an expected consequence of BGC if the rate of BGC is correlated with that of recombination. There is also the possibility of selection for GC bases, but, following arguments from Sharp et al. (1995), Galtier et al. (2001) and Marais (2003) argue that this type of selection seems unlikely for non-coding sequence. They argue that natural selection controlling GC content at every nucleotide in the genome would require very high selection coefficients, especially in species with low effective population size, at every position of GC-rich regions, which seems unlikely given the lack of correlation with gene expression. Haddrill and Charlesworth (2008) also observed a positive correlation between GC content and the proportion of GC to AT vs. AT to GC polymorphisms, indicating that GC-rich sequences have a stronger selection or biased gene conversion favouring GC variants.

1.3.4 Conclusion

From these recent studies, it seems clear that mutations arising in non-coding sequences are selectively constrained in *Drosophila*. Indeed, using different methods and types of sequence data, the results generally imply the presence of both positive and negative selection on non-coding DNA. While genome-wide comparisons can span very large amounts of sequences and are not biased towards particular classes of sequences, polymorphism data allows more powerful tests of natural selection. Furthermore, it is possible that some functions are conserved between species with little overall sequence similarity, but this only means that focusing on conserved sequences would slightly underestimate functional non-coding sequence.

For future research in order to confirm these results for the whole genus, it would be useful to study different species, and also localise constrained sequences more

precisely. Taking into account insertions and deletions could also provide more information on the processes that conserve non-coding DNA.

The concept of deleterious effects of mutations has been mainly used for mutations resulting in a change in the amino-acid sequences in coding regions. However, current results suggest that many deleterious mutations may occur in non-coding DNA. Another conclusion arising from these studies is the existence of many possibly correlated factors (GC content, recombination rate, etc...) that need to be teased apart. Finally, as we are only just starting to study non-coding DNA, experiments, such as removing blocks of constrained non-coding sequence in laboratory flies, could be designed to gain a better understanding of the function of these constrained sequences.

1.4 Aims of this study

This thesis explores different methods to detect natural selection in DNA sequences, specifically in non-coding DNA sequences, and parameters that can affect DNA sequences as well. Five different studies have been carried out, all of which address questions related to the effects of natural selection on DNA sequences.

In Chapter 2, I obtained sequences for a large number of genes in *D. miranda* from BAC sequences, and compared these with sequences from its close relative, *D. pseudoobscura*. Divergence studies can point to constrained regions where purifying selection has kept divergence levels low. As in previous studies in *D. melanogaster*, I found a negative relationship between intron length and intron divergence, suggesting that longer introns are under selective constraint. I also find a negative correlation between the rate of non-synonymous substitutions and codon usage bias, suggesting that fast-evolving genes have a lower codon usage bias, consistent with strong positive selection interfering with weak selection for codon usage.

Secondly, in Chapter 3, I gathered polymorphism data for a smaller number of genes in *D. americana* in order to distinguish between positive and negative selection using methods that require polymorphism and divergence. I fail to detect a significant

relationship between intron length and divergence or polymorphism. Surprisingly, the direction of this relationship seems to be the opposite of previous findings, with larger introns being more diverged than smaller introns. First introns show lower polymorphism and divergence than non-first introns, suggesting that they may be more constrained, although the difference is not significant.

Using the same *D. americana* dataset, I then focused, in Chapter 4, on insertions and deletions to test the hypothesis that insertions are favoured to compensate for the deletion bias in *Drosophila*. I used a maximum-likelihood method that takes into account demographic history, in this case a recent population expansion and then calculates the selection coefficients. Although the results were not significant, the values suggest weak positive selection acting on insertions, as expected.

I then tested in Chapter 5 for selection on GC to AT polymorphisms in the *D. americana* intron dataset. We expect to observe as many GC to AT as AT to GC segregating mutations and that they should have similar mean frequencies if no selection is acting. I find evidence for a preference for GC in my dataset. I also investigate codon usage bias using preferred and unpreferred codons changes and frequencies and I find evidence for selection for codon usage bias. I then studied the effect of recombination rates on these patterns by looking at the difference between rates in exons and rates in introns. Using LDhat on the *D. americana* dataset, I find that recombination estimates are not significantly different between introns and coding sequences, which is of significance in relation to interpretations of differences in the apparent strength of selection on non-coding and synonymous sites.

Finally, in Chapter 6, I looked at gene expression as a factor that can affect natural selection. I used gene expression data from seven *Drosophila* species to test the hypothesis that genes on the 4th chromosome or Muller element F, which has low crossing-over, have higher gene expression than genes on other chromosomes as previously found. I find that microarray data yields the opposite result to the EST data, either suggesting that gene expression is actually lower on Muller element F or that there are biases in this method.

2 Patterns of DNA sequence divergence between *Drosophila miranda* and *D. pseudoobscura*.

The work described in this Chapter has been recently published (Marion de Procé et al. 2009).

Contributing authors:

- I performed the alignments, data extraction and data analysis and I wrote the manuscript.
- D. Halligan helped with Perl and R scripts for data extraction and analyses and gave comments on the manuscript.
- P. Keightley gave comments on the manuscript.
- B. Charlesworth advised on the project and helped write the manuscript.

2.1 Introduction

Several studies have shown that non-coding DNA is more highly constrained on average than synonymous sites between *Drosophila melanogaster* and its close relative *D. simulans* (Bergman and Kreitman 2001; Halligan et al. 2004; Andolfatto 2005; Haddrill et al. 2005; Marais et al. 2005; Halligan and Keightley 2006; Casillas et al. 2007; Haddrill et al. 2008a), suggesting that much non-protein-coding DNA in *Drosophila* is functional. Similar studies on mammalian genomes suggest that a smaller fraction of non-coding DNA is functional (Birney et al. 2007). One of the more surprising findings in *Drosophila* has been a significant negative correlation between intron length and intron divergence between *D. melanogaster* and *D. simulans* (Parsch 2003; Marais et al. 2005; Haddrill et al. 2005). These findings were confirmed by a whole-genome study of *D. melanogaster* and *D. simulans*, which showed that the level of selective constraint is positively correlated with intronic as well as intergenic sequence length (Halligan and

Keightley 2006). Another observation is that first introns are more conserved in mammals, as they contain more regulatory elements (Majewski and Ott 2002; Keightley and Gaffney 2003; Chamary and Hurst 2004), and Duret (2001) and Marais et al. (2005) suggest that it could also be the case in *Drosophila*. First introns are also longer than other introns in mammals (Smith 1988) and in *Drosophila* (Maroni 1994; Duret 2001; Marais et al. 2005; Bradnam and Korf 2008). The relationship between intron divergence and intron length could thus be affected by the position of introns, although Haddrill et al. (2005) found that mean divergence did not differ between first and non-first introns within short and long intron size categories. Levels and patterns of constraint on intergenic sequences appear to be broadly similar to those on long introns (Bergman and Kreitman 2001; Andolfatto 2005; Halligan and Keightley 2006).

It is important to determine whether these patterns of sequence evolution apply more generally. This can be done using comparisons of species that are sufficiently closely related that their non-coding sequences can reliably be aligned, but are distant enough that there is some power to detect patterns of divergence. Unfortunately, the 12 species of *Drosophila* that have been sequenced (Clark et al. 2007) are far from ideal for this purpose. For this reason, we have chosen to compare the close relatives *D. miranda* and *D. pseudoobscura*. The latter is one of the 12 sequenced Drosophila genomes (Richards et al. 2005), and its approximate divergence time from *D. miranda* is 2 My (Barrio et al. 1992), with an average divergence at fourfold synonymous sites of 3.6% (Bachtrog and Andolfatto 2006). A high rate of chromosomal rearrangements has been found between these two species (Bartolomé and Charlesworth 2006a). Recently, long introns were shown to be less diverged than short introns between *D. pseudoobscura* and *D. miranda* (Bachtrog and Andolfatto 2006), consistent with the results for *D. melanogaster* and *D. simulans*.

Other patterns that can be explored with this comparison are as follows. It has been shown in comparisons of *D. melanogaster* with its relatives that genes with high levels of nonsynonymous divergence have lower codon usage bias, possibly caused by selective interference from positively selected nonsynonymous mutations, or because of a general reduction in selective constraints on these genes (Betancourt and Presgraves

2002; Marais et al. 2004; Bierne and Eyre-Walker 2006; Andolfatto 2007; Bachtrog 2008). However, using a molecular-level evolutionary simulation, Drummond and Wilke (2008) showed that this relationship might be explained by selection against the toxicity of misfolded proteins induced by mistranslation. *D. miranda* coding sequences have been shown to be under weak selection for codon usage (Bartolomé and Charlesworth 2006b; Bachtrog 2007). Gene expression levels are highly correlated with optimal codon usage (Duret and Mouchiroud 1999), which has been interpreted as evidence for selection leading to highly expressed genes having optimal codons for more efficient or accurate translation. It is therefore important to correct for the effects of gene expression on patterns of codon usage and sequence divergence. This has not always been done, although Marais et al. (2004) found that correcting for gene expression levels estimated from microarray data did not alter the negative correlation between codon usage and divergence between *D. melanogaster* and *D. simulans*. Since expression data are available for *D. pseudoobscura*, the *miranda-pseudoobscura* comparison offers an opportunity to examine the questions raised by the studies of *D. melanogaster* in an independent contrast of two species.

The aim of this chapter was to analyze alignments of the sequences of a set of *D. miranda* BAC clones and the corresponding parts of the *D. pseudoobscura* genome sequences, and to determine whether the relationships described above hold for these two species. An advantage of using material from BAC clones is that they represent an unbiased sample of genes, whereas studies that use primers designed for coding or non-coding sequences, such as studies on *D. miranda* (Bachtrog 2008; Bachtrog and Andolfatto 2006), may be biased towards more conserved sequences.

2.2 Material and methods

BAC libraries for *D. miranda* were created by Dr. Xulio Maside (University of Edinburgh) and the Children's Hospital Oakland Research Institute (CHORI) (Bachtrog et al. 2008) (Table 1), and sequenced at the Wellcome Trust Sanger Institute. We aligned these sequences to the corresponding sequences of the *D. pseudoobscura* genome, for the purpose of studying sequence divergence between the two species.

2.2.1 BAC sequencing

Fly material.

As described in Bachtrog et al. (2008), high molecular weight DNA suitable for creating BAC libraries was isolated by Xulio Maside from adult males from a *D. miranda* isofemale line (MSH22: Yi and Charlesworth 2000), which has been maintained in laboratory culture for more than 10 years.

BAC library

A *D. miranda* BAC library was produced by the CHORI in a pTARBAC6 vector. The library was tested with the following amplicons by Dr. Mark Dorris in Edinburgh, in order to localize a subset of BAC sequences within the *D. miranda* genome: 1, *DdC*; 2, d*pp*; 3, *Eno*; 4, *Gpdh*; 5, *bcd*; 6, *Gld*; 7, *hb*; 8, *Rp49*; 9, *Est-5B*; 10, *Gapdh2*; 11, *swallow*; 12, *sesB*. The sequences for these probes were provided by Dr Carolina Bartolomé, as described in Bartolomé et al. (2005). Positive colonies established by individual amplification using the above amplicons were picked onto agar stabs and sequenced at the Wellcome Trust Sanger Institute.

There is no full sequence for BAC 10 due to its containing a high number of repeats, BAC 9 contains a 120 bp region that could not be sequenced because it is surrounded by long runs of G's and C's; and BAC 6 is made of two contigs separated by a 1225 bp gap, possibly with a high A/T content. The repeats in BAC10 may be associated with the transposition of this region from XL to XR in *D. pseudoobscura* (Bartolomé and Charlesworth 2006a).

Characterization of the BAC clones

I used 12 BAC sequences (~200 kb each) from chromosomes 2, 4, XL and XR of *D. miranda* (Table 2.1). I located orthologous sequences by BLASTing 200 bp from every 1200 bp of the *D. miranda* BAC sequences against the repeat-masked *D. pseudoobscura* genome (release dp3 from UCSC genome browser; Richards et al. 2005). In the masked version of the genome, repeats from RepeatMasker and Tandem Repeats Finder are masked. I plotted the location on dp3 contigs against the location on the BAC sequence to identify sections that were co-linear, and all contiguous BLAST hits were grouped into fragments. The homologous sequences were aligned with MAVID (Bray and Pachter 2004). I extracted the coding, intronic and intergenic alignments for 192 genes from the large BAC alignments by mapping the *D. pseudoobscura* annotation on to the alignments (Appendix 2.1). Some genes were found to overlap with each other in two ways: either a whole gene was included in the large intron of another gene, or two genes were overlapping for most of the sequence. Analyses were done excluding these overlapping genes to avoid data duplication, and to ensure that all sequence identified as intergenic or intronic were completely non-coding. Introns and intergenic sequences were then realigned using MCALIGN2 (Wang et al. 2006), using an insertion-deletion frequency model previously defined for *Drosophila* intronic DNA (Keightley and Johnson 2004). Sequences were deposited into GenBank under accession numbers FJ821025-FJ821035.

Table 2.1: Identification of *D. pseudoobscura* orthologs.

Clone (GenBank ID)	Initial probe	Gene[a] ID	*D. mir* location[b]	Part[c]	Length	No. of genes[d]	Blasts *D. pse*[e]
BAC1 (FJ821025)	*Ddc*	GA10503	4 (68F)	1	99,600	7	chr4_group4
				2	31,174	8	chrU[f]
				3	35,771	1	chr4_group4
				4	38,529	8	chr4_group1
BAC2 (FJ821028)	*dpp*	GA22099	4 (63C)	1	205,597	9	chr4_group4
				2	13,000	0	chr4_group4
				3	5,000	0	chr4_group3
				4	7,000	0	chr4_group3
BAC3 (FJ821029)	*Eno*	GA14598	4 (61C)	1	160,000	11	chr4_group3
				2	22,000	1	chr4_group3
BAC4 (FJ821030)	*Gpdh*	GA21498	4 (66C)	1	199,514	7	chr4_group1
				2	15,000	0	chr4_group4
BAC5 (FJ821031)	*bcd*	GA10255	2 (48A)	1	189,519	11	chr2
BAC6 (FJ821032)	*Gld*	GA11047	2 (53C)	1	187,482	18	chr2
BAC7 (FJ821033)	*hb*	GA22036	2 (53B)	1	189,926	31	chr2
BAC8 (FJ821034)	*RpL32* (*Rp49*)	GA20704	2 (35F-36A)	1	229,949	27	chr2
				2	5,000	0	chr3
BAC9 (FJ821035)	*Est-5B*	GA14349	XR (17F)	1	184,788	19	chrXR_group6
				2	11,000	0	chrU
BAC10	*Gapdh2*	GA21397	XL (2D)	NA	NA	0	NA
BAC11 (FJ821026)	*sesB*	GA14229	XL (5B)	1	211,219	13	chrXL_group1e
BAC12 (FJ821027)	*swallow*	GA17446	XL (8A)	1	183,049	14	chrXL_group1a
				2	70,000	0	chrXL_group1a

[a] known gene in each BAC sequence; [b] expected location in *D. miranda*; [c] when a BAC sequence has several parts, it means that there were groups of BLAST hits that were not contiguous, the BLAST hits are contiguous and collinear with the *D. pseudoobscura* genome in each individual part; [d] genes that were included in the analyses; [e] location of BLAST hits of the BAC sequence on the *D. pseudoobscura* genome; [f] chrU is a concatenation of unplaced contigs.

2.2.2 Coding sequences

Sequences with internal stop codons in *D. miranda* were excluded, since these may represent sequencing or alignment errors, or genes that have lost their function in *D. miranda*. Coding sequences were checked against the Flybase coding sequences of *D. pseudoobscura* to ensure that the annotations agreed. After rejecting two genes that had internal stop codons, 5 incomplete coding sequences and 3 overlapping genes, 182 coding sequences remained in the final dataset.

For the analyses of coding sequences, we estimated the frequency of optimal codons (F_{op}: determined using Codonw (http://codonw.sourceforge.net) from the *D. pseudoobscura* preferred codon table (Vicario et al. 2007), and d_N/d_S (PAML: Yang 1997) and K_a/K_s (using the method of Comeron 1995 implemented in Gestimator: libsequence C++ library, Thornton 2003) ratios. K_a and d_N measure the rate of non-synonymous substitutions, and K_s and d_S estimate the rate of synonymous substitutions. Gene length was calculated from the start of the first codon to the end of the last codon, leaving out introns and UTRs. We also used gene expression as a covariate, so we used expression data for *D. pseudoobscura* from the GEO database. This database was generated by Zhang et al. (2007), who recently performed microarray experiments to investigate sex-biased expression of orthologues and species-restricted genes in *Drosophila* (data accessible at NCBI GEO database (Edgar et al. 2002), accession GSE6640). We used the log2 transformed signal intensities after VSN (Variance Stabilization Normalization) transformation (Huber et al. 2002), which were available for 172 genes in our dataset. These data were available for five males and four females, so we calculated the weighted average of these values for each gene.

2.2.3 Introns

In order to avoid regions where constraint due to splicing mechanisms is already documented (Halligan and Keightley 2006), we removed 8bp from the 5' end and 30bp from the 3' end of introns, which correspond to the splice sites of introns. Leaving these sites in the intron sequences would weaken any correlation with length, since these always contribute the same number of bases to any other intron, and proportionally less

to longer introns. We discarded 11 introns that were overlapping with coding sequence from other genes, so that the final dataset comprises 406 introns. We estimated GC content, intron length, and divergence between *D. miranda* and *D. pseudoobscura* using the Jukes-Cantor correction (Jukes and Cantor 1969). Since the mean silent divergence between these species is low (~3.6%), this should be sufficiently accurate for the purpose of this study. We split introns into three length categories: short introns between 51 bp and 80 bp (284 introns), long introns between 81 bp and 500 bp (52 introns), and very long introns over 500 bp (70 introns). This meant that we omitted 20 introns of less than 51 bp.

In order to determine intron lengths in the *D. pseudoobscura* genome as a whole, we also extracted the start and end positions of introns as well as their position in the gene from the DroSpeGe database. There are 5986 first introns and 12961 non-first introns in this dataset.

2.2.4 Intergenic sequences

We defined intergenic sequences as those regions between the ends and the starts of coding sequences. We discarded 5 intergenic sequences that overlapped with coding sequence, giving 167 intergenic sequences in the final dataset. UTRs are unannotated in the *D. pseudoobscura* genome, so we analyzed separately the start, middle and end of intergenic sequences. Evolutionary divergences were obtained for the whole sequence, as well as for the edges and the centre of the sequence, in order to detect the potential effects of UTRs at the edges of these intergenic sequences. Halligan and Keightley (2006) found the average length of 5' UTRs and 3' UTRs in *D. melanogaster* to be 148 and 280bp long respectively, so we used these mean values to define the edges of intergenic sequences. Intergenic sequences were all longer than 80bp, so they were split into only two length categories, long and very long, as explained in the introns section.

2.2.5 Analyses

Analyses were done using the R statistical package (http://www.r-project.org/). 95% confidence intervals for partial Pearson correlation coefficients were obtained by bootstrapping 1,000 times by sequence for the coding sequence and intergenic sequence datasets. For the intron dataset, however, an ANOVA showed significant variation for intronic divergence among genes (p=0.009), indicating that divergence values for introns within a same gene are not independent. Thus, for this dataset we bootstrapped 1,000 times by gene. Wilcoxon two-sample tests were performed to compare means between categories of sequences (e.g. long versus short and X-linked versus autosomal.). Paired bootstrap tests were performed to test the difference in mean divergence between sequence categories and 95% confidence intervals were calculated by bootstrapping 100,000 times by gene.

2.3 Results

2.3.1 Relationship between non-coding sequence GC content and divergence

Haddrill et al. (2005) found a significantly negative correlation between intron divergence and GC content, which could influence the relationship between intron length and divergence. Here, we found that intron divergence and GC content are negatively but not significantly correlated (Figure 2.1 for *D. miranda*) (*D. pse*: Spearman r_S = -0.091, p = 0.068; *D. mir*: Spearman r_S = -0.084, p = 0.093). After accounting for intron length, the partial Pearson correlation coefficient for divergence and GC content is r = -0.051, 95% C.I. = [-0.151; 0.047]. Although it is not significant, the relationship is in the same direction as found by Haddrill et al. (2005). There is a negative but non-significant correlation between intergenic sequence divergence and GC content; the partial correlation coefficient, after accounting for intergenic sequence length, is Pearson r = -0.012, 95% C.I. = [-0.182; 0.159].

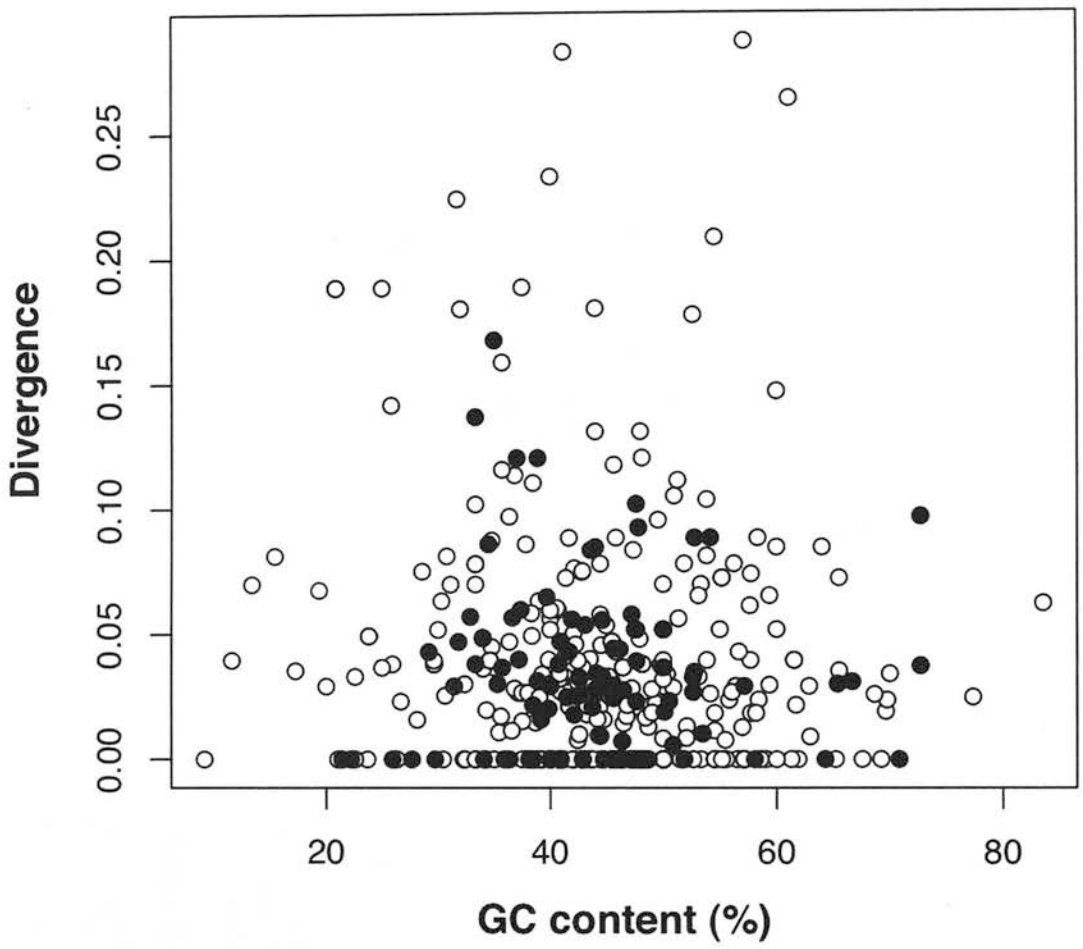

Figure 2.1: Plot of intron divergence against GC content in *D. miranda*. Solid circles represent first introns and open circles represent non-first introns.

2.3.2 Relationship between non-coding sequence length and divergence

The first relationship that we investigated in this context was the correlation between non-coding sequence length and divergence, since Haddrill et al. (2005) and Halligan and Keightley (2006), and Bachtrog and Andolfatto (2006) all found a negative correlation between intron length and divergence using *D. melanogaster* and *D. simulans*, and *D. pseudoobscura* and *D. miranda*, respectively. This was also observed in our data, even after correcting for GC content (Figure 2.2) (Pearson r = -0.064; 95% bootstrap by gene C.I. = [-0.098; -0.039]). Using the same method but accounting for gene expression this time, the correlation coefficient is Pearson r = -0.057; bootstrap by gene 95% C.I. = [-0.096; -0.037]). Gene expression is positively correlated with intron length (Spearman r_S

= 0.262, p = 0.002), as Marais et al. (2005) found for first introns in *D. melanogaster*, but there is no significant correlation with intron divergence (Spearman r_S = 0.006, p = 0.909). Accounting for both GC content and gene expression, the correlation coefficient between intron length and divergence is Pearson r = -0.057; 95% bootstrap by gene C.I. = [-0.094; -0.034], so the negative correlation coefficient is still significant when both variables are accounted for. These correlation coefficients are lower than those for the *melanogaster-simulans* comparison, possibly reflecting the smaller levels of divergence in our case, with correspondingly more noise in the estimates relative to the mean divergence.

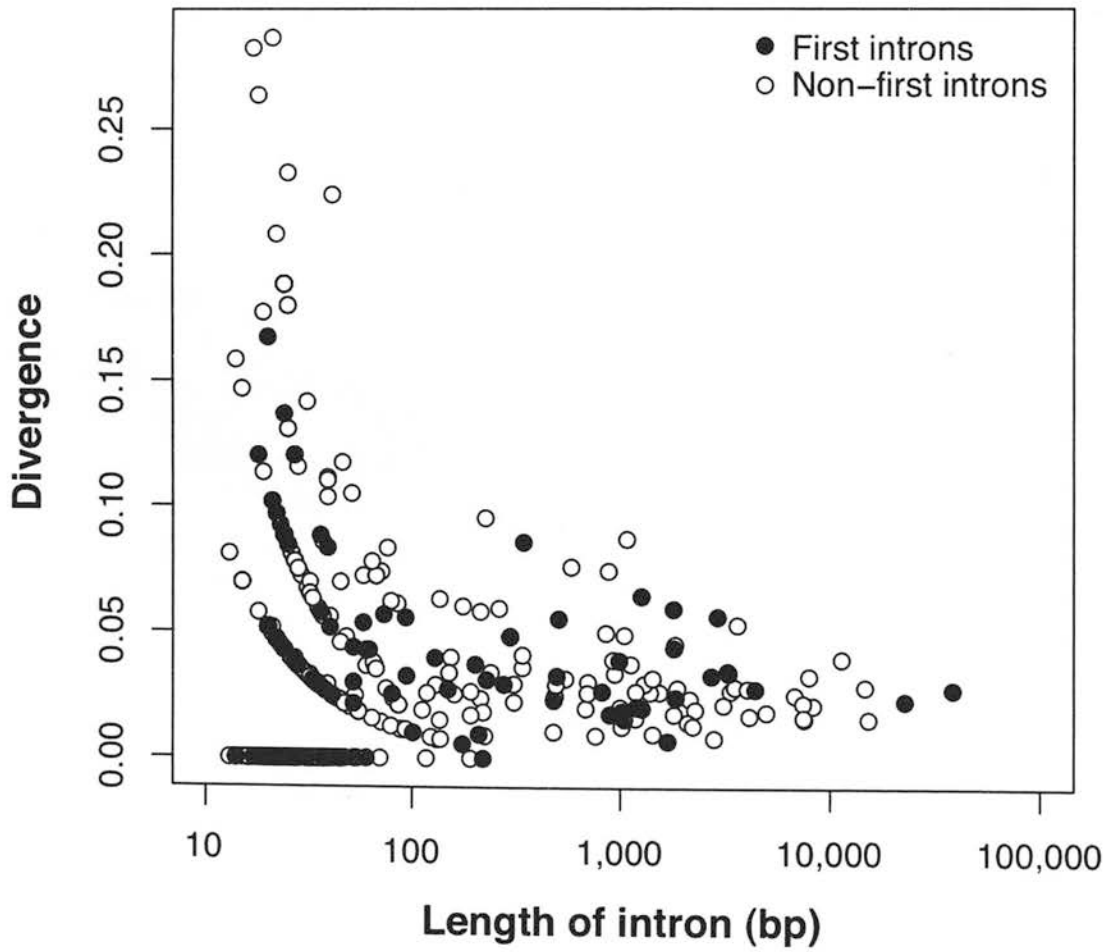

Figure 2.2: Plot of intron divergence against intron length on a log scale. Solid circles represent first introns and open circles represent non-first introns.

The correlation coefficient between evolutionary divergence and intergenic sequence length, after accounting for GC content, is Pearson $r = 0.114$; 95% C.I. = [-0.010; 0.244] (Figure 2.3). The correlation is non-significantly different from zero, and the estimate is positive, and plausibly could be at most slightly negative. Halligan and Keightley (2006) found a significantly negative correlation between divergence and intergenic sequence length for their genome-wide *melanogaster-simulans* comparison, so that our result could entirely be due to the small number of intergenic sequences surveyed in this study. There is some variation in constraint levels in intergenic sequences, since divergence is smaller at the edges of intergenic sequences (mean divergence: 0.027) than in their middle (mean divergence: 0.031) (one-sided Wilcoxon test, p-value = 0.00357). This test suggests that the edges of intergenic sequences are more strongly constrained than the middle, possibly due to the presence of promoters or UTRs. Bachtrog and Andolfatto (2006) also found high levels of constraint (~30%) between *D. pseudoobscura* and *D. miranda* in intergenic sequences, a value that is likely to be due to the presence of UTRs. They measured constraint as the percentage reduction below the divergence at synonymous sites.

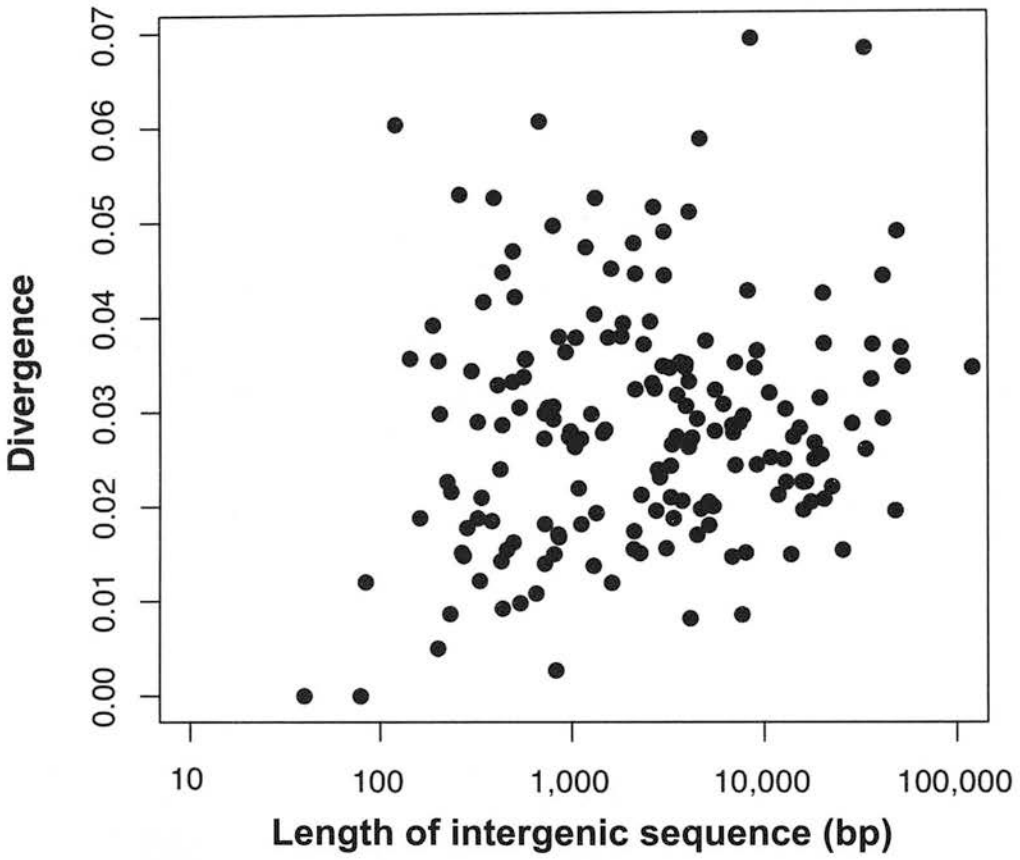

Figure 2.3: Plot of intergenic sequences divergence against intergenic sequence length on a log scale.

Figure 2.4 shows the mean divergence levels of different categories of coding and non-coding sites. As expected, non-degenerate sites have a much lower divergence than all other classes (Wilcoxon tests, $p \leq 0.0027$; $p \leq 2\times10^{-5}$ from all paired bootstrap differences between the non-degenerate sites divergence and other divergences). Synonymous sites are more diverged than long and very long intergenic and non-first intronic sequences (Wilcoxon test, $p \leq 0.0172$; $p \leq 2\times10^{-5}$ from all paired bootstrap differences between the synonymous sites divergence and intron divergences), except for first introns, but the estimate of the mean divergence for first introns has large standard errors. Short first introns have a lower divergence than short non-first introns but the difference is not statistically significant (Wilcoxon test, $p = 0.285$). The difference in length between first and non-first introns is non-significant (Wilcoxon test, $p = 0.18$), with first introns being shorter than long introns, which contrasts with previous

observations in *D. melanogaster* (Maroni 1994; Duret 2001; Bradnam and Korf 2008) and also with the whole *D. pseudoobscura* genome data, which shows that first introns are significantly longer than non-first introns (Wilcoxon test, $p = 1.03 \times 10^{-08}$). It is thus likely to be only a sample effect in our dataset. Short introns show higher divergence than long and very long introns, significantly so for very long introns using the paired bootstrap test (long: W = 7995, p = 0.34; very long: W = 10463, p = 0.49; p = 0.02 from all paired bootstrap differences between short intron divergence and very long intron divergence). The ratio of the mean divergence for all long introns to the mean divergence for short introns is 0.652, which is similar to 0.636, the ratio found in the *melanogaster-simulans* comparison (Haddrill et al. 2005). Divergence in short introns is not significantly different from that for synonymous sites (Wilcoxon test, p > 0.09).

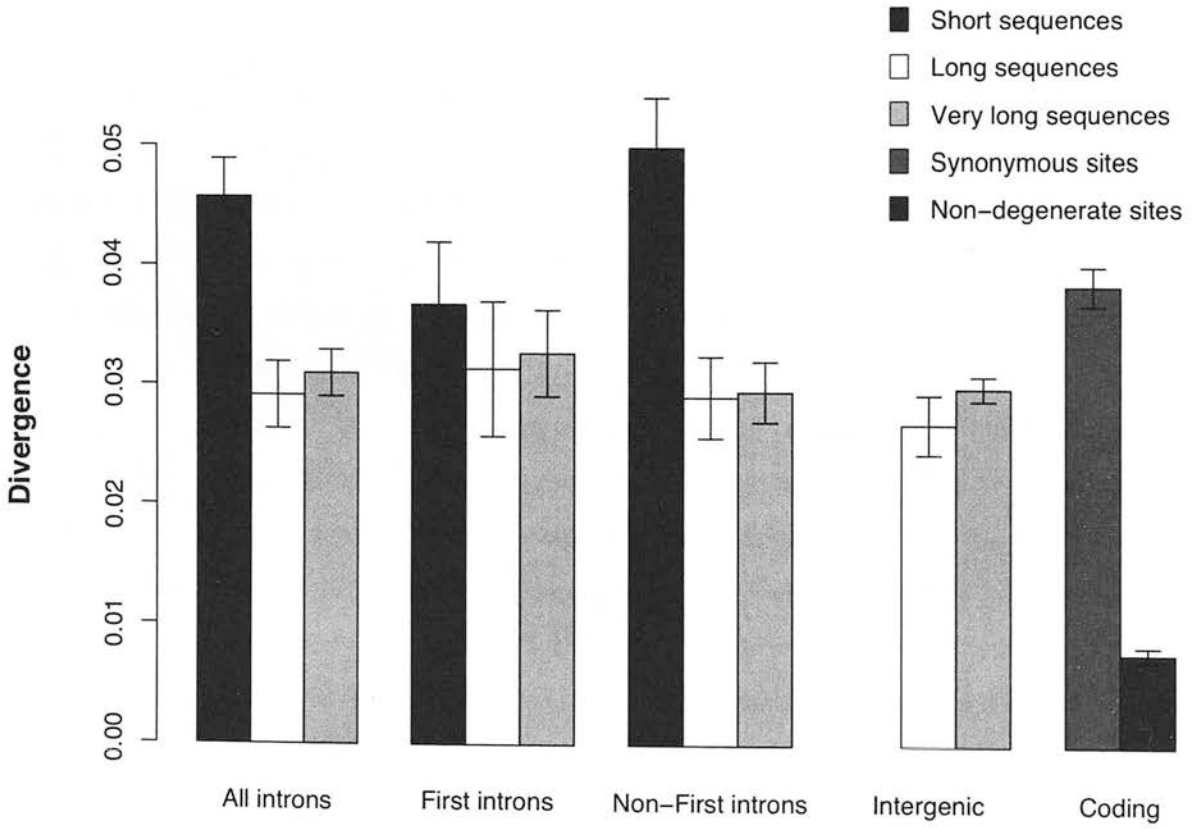

Figure 2.4: Barplot of mean divergence (+/- S.E.) for different classes of sites. Rates were calculated using the Jukes-Cantor multiple hit correction. Short sequences are under 80 bp, long sequences are between 80 bp and 500 bp and very long sequences are over 500 bp.

2.3.3 Comparison of X-linked versus autosomal loci

Differences between X-linked and autosomal loci have been observed in previous studies, that suggest faster evolution and higher codon usage bias for X-linked loci (Charlesworth et al. 1987; Singh et al. 2005b, Larracuente et al. 2008), so we looked for the same patterns in our dataset. To account for the effect of gene length and gene expression, we separated the data for coding sequences and introns for X-linked and autosomal loci into two length categories (with the cutoff gene length set at the median 1100bp) and separately into two expression level categories (with the cutoff expression value set at the median 9.3).

When using the dataset before separation into categories, X-linked coding sequences have a higher GC content, particularly GC3 content, and higher codon usage indices than autosomal coding sequences (Wilcoxon tests, $p < 0.004$) (Figure 2.5). For each category considered separately, these differences remain significant (Wilcoxon tests, $p < 0.002$), supporting previous findings. However, none of these differences are significant in the high expression category of genes (Wilcoxon tests: $p > 0.1$), and the difference in GC content is only near significance in the short genes category. All these results are the same in *D. pseudoobscura* and *D. miranda*. There is no significant difference in the rate of synonymous or non-synonymous substitutions between X-linked and autosomal coding sequences.

In the non-coding sequences, before separation into categories, X-linked introns have the same divergence as autosomal introns (Wilcoxon test: $p = 0.06$), but a higher GC content (Wilcoxon tests: $p < 0.017$) and intergenic sequences have the same GC content and rate of substitutions on the X and autosomes (Wilcoxon tests: $p > 0.06$). For each category of introns considered separately, the difference in rate of substitutions remains non-significant. The difference in GC content between X-linked and autosomal introns is more highly significant in long genes (Wilcoxon tests, $p < 0.005$) but becomes non-significant in other intron categories (Wilcoxon tests, $p > 0.06$).

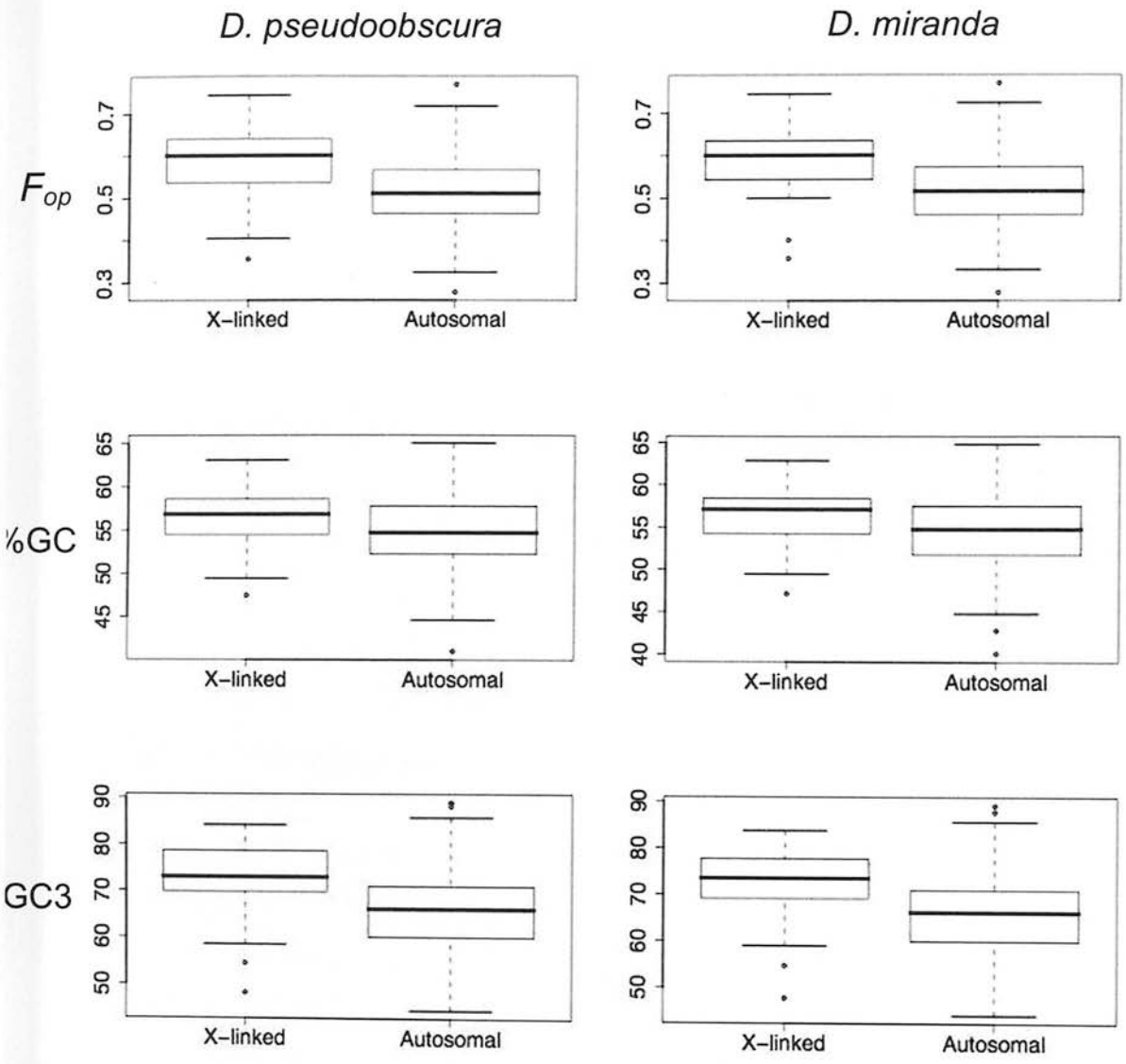

Figure 2.5: Box plots for *Fop*, GC and GC3 content of X-linked versus autosomal loci for *D. pseudoobscura* and *D. miranda*.

2.3.4 Relationship between codon usage (F_{op}) and d_N

A negative correlation has previously been observed between codon usage (*Fop*: frequency of preferred codons; Ikemura 1981) and the rate of non-synonymous substitutions (measured by d_N or K_a) between *D. melanogaster* and *D. simulans* (Betancourt and Presgraves 2002) and between *D. pseudoobscura* and *D. miranda* (Bachtrog 2008). This suggests that genes with fast-evolving protein sequences have lower codon usage bias, which implies that selection for codon usage is less effective in

such genes. However, gene expression might affect this relationship, because highly expressed genes are usually more conserved. This should therefore be corrected for in such an analysis. Marais et al. (2004) have indeed shown that d_N and expression level are negatively correlated in a dataset of 630 orthologous sequence pairs from *D. melanogaster* and *D. yakuba*, although this correlation is weaker than that between d_N and F_{op}. Furthermore, Andolfatto (2007) suggested that the silent substitution rate should be accounted for when calculating the correlation between codon usage and the rate of non-synonymous substitutions because of the positive correlation between K_a and K_s (Comeron and Aguadé 1996). Our results also show a significant positive correlation between K_a and K_s (Spearman r_S = 0.409, p = 1.55×10^{-8}). Codon usage bias decreases with gene length in *D. melanogaster* (Duret and Mouchiroud 1999) and long proteins are expected to be disadvantageous (Moriyama and Powell 1998), so gene length is another factor to correct for in this analysis.

F_{op} and d_N are significantly negatively correlated when correcting for d_S (Figure 2.6) (*D. pse*: Pearson r = -0.419, 95% bootstrap C.I. = [-0.528; -0.301]; *D. mir*: Pearson r = -0.415, 95% bootstrap C.I. = [-0.531; -0.304]). After controlling for gene length, gene expression, and d_S, there is still a significant negative correlation between *Fop* and d_N (*D. pse*: Pearson r = -0.278, 95% bootstrap C.I. = [-0.468; -0.085]; *D. mir*: Pearson r = -0.313, 95% bootstrap C.I. = [-0.478; -0.139]). Thus, as found in previous studies (Bachtrog 2008), codon usage in *D. pseudoobscura* and *D. miranda* appears to decrease as the non-synonymous substitution rate increases, even after controlling for potential effects of gene length, gene expression and d_S.

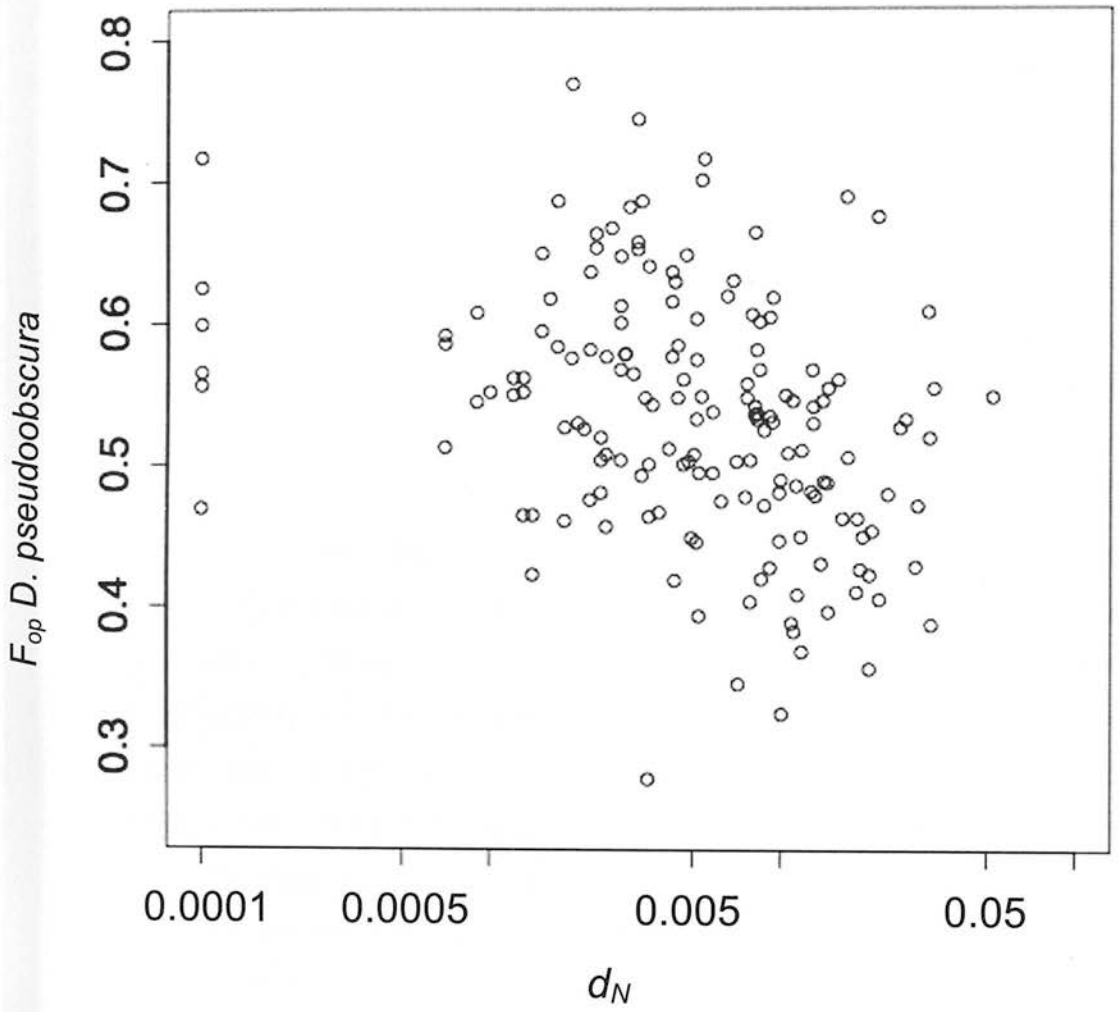

Figure 2.6: Plot of *Fop* for *D. pseudoobscura* against d_N on a log scale.

F_{op} and d_S are significantly positively correlated after correcting for d_N, gene length and gene expression (*D. pse*: Pearson r = 0.457; 95% bootstrap C.I. = [0.113; 0.656]; *D. mir*: Pearson r = 0.481; 95% bootstrap C.I. = [0.150; 0.682]). This effect is probably due to the influence of base composition on d_S (Bierne and Eyre-Walker 2003), and becomes non-significant when using K_a and K_s instead of d_N and d_S (*D. pse*: Pearson r = 0.264; 95% bootstrap C.I. = [-0.044; 0.516]; *D. mir*: Pearson r = 0.299; 95% bootstrap C.I. = [-0.015; 0.529]).

Another finding is the positive correlation between F_{op} and coding sequence length after correcting for gene expression (*D. pse*: Pearson $r = 0.182$, 95% bootstrap C.I. = [-0.073; 0.435]), which contradicts results from Duret and Mouchiroud (1999), but is in agreement with the model of selection on translational accuracy (Drummond and Wilke 2008), according to which codon usage bias should be higher in genes encoding long proteins. However, this correlation is not significant, even after correcting for d_N (Pearson $r = 0.213$, 95% bootstrap C.I. = [-0.045; 0.423]). We also found a highly significant positive correlation between gene expression and codon usage after correcting for gene length and K_s (Pearson $r = 0.513$, 95% bootstrap C.I. = [0.291; 0.656]), consistent with previous results (Shields et al. 1988; Moriyama and Powell 1998; Duret and Mouchiroud 1999) and the theory that highly expressed genes experience stronger selection on translational accuracy (Moriyama and Powell 1998; Drummond and Wilke, 2008).

We also found that d_N and gene expression are negatively correlated after accounting for gene length, d_S and GC content (Pearson $r = -0.237$, 95% bootstrap C.I. = [-0.432; -0.050]), consistent with the results from Marais et al. (2004) and Subramanian and Kumar (2004), who found that highly expressed proteins evolve slowly in flies and in vertebrates, respectively. d_S and gene expression are also negatively correlated, after accounting for d_N, gene length and GC (Pearson $r = -0.272$, 95% bootstrap C.I. = [-0.443; -0.089]). These two relationships also agree with findings of Drummond and Wilke (2008), supporting their hypothesis of selection against misfolded proteins.

2.4 Discussion

Our study sheds new light on some important aspects of sequence evolution in *Drosophila*, by using sequence comparisons between a set of more or less randomly chosen loci for a pair of closely related species in the *obscura* group, *D. pseudoobscura* and *D. miranda*. These species are sufficiently distantly related that there is some power to detect patterns of sequence evolution revealed by earlier studies, predominantly of the *melanogaster* group, unlike comparisons between *D. pseudoobscura* and *D. persimilis*, for which whole genome sequences are available but which are almost undistinguishable at the sequence level (Clark et al. 2007). They are sufficiently close that alignments of non-coding sequences can be reliably performed, making them a useful tool for our purpose, in addition to the intrinsic importance of these species for other problems in evolutionary genetics, such as Y chromosome evolution (Bachtrog et al. 2008).

There are several limitations to our study. First, the distribution of intron lengths is highly skewed, with many more short introns than long introns, which reduces the power of correlation-based tests. Second, *D. miranda* and *D. pseudoobscura* are closely related and show signs of divergence caused by different ancestral polymorphic variants having becoming fixed independently in the two species, rather than by fixation of new mutations (Bartolomé et al. 2005). In addition, by using only one sequence from each species, polymorphism and divergence are confounded, which will result in overestimates of true divergence levels and underestimation of the effects of purifying selection, since this has less influence on polymorphism than divergence (Akashi 1995; Charlesworth 1994). Finally, a variable that has not been taken into account in this study, due to a lack of detailed information, is the recombination rate. This has been shown to affect GC content (e.g. the positive correlation between recombination rate and GC content in large introns and intergenic regions (Marais et al. 2003)) and the efficacy of selection strength (Comeron et al. 2008). Too little reliable information is currently available on this variable.

2.4.1 Non-coding sequence length and divergence

There is a negative correlation between intron length and intron divergence between *D. melanogaster* and *D. simulans* (Haddrill et al. 2005; Halligan and Keightley 2006), and long introns were found to be less constrained than short introns in a comparison of *D. pseudoobscura* and *D. miranda* (Bachtrog and Andolfatto 2006). The present study confirms this observation. Studies using polymorphism data in addition to divergence data have also shown that long intron sequences are under purifying natural selection (Andolfatto 2005; Casillas et al. 2007; Haddrill et al. 2008a), so this does not simply reflect differences in mutation rates between long and short introns.

Several possible explanations for the conservation of long non-coding sequences have been suggested. First, longer sequences may contain more *cis*-regulatory elements (Bergman et al. 2002; Emberly et al. 2003; Sironi et al. 2005). Another complementary explanation is based on the observation of a general mutational bias in favour of deletions in *Drosophila* (Petrov 2002). If a sequence is functionally constrained, deletion bias will be countered by selection and the sequence will be longer than non-conserved sequences. To explain the persistence of short poorly conserved sequences, we might argue that it is harder to fix deletions in short introns, because deletions are more likely to affect adjacent coding sequence. Furthermore, the need for a minimum intron size for correct splicing (Mount et al. 1992) suggests that short introns would merely be spacers between exons. This suggests that most of the non-coding DNA in the *Drosophila* genus has been conserved because it has a function.

First introns are longer (Duret 2001; Bradnam and Korf 2008) and possibly contain more regulatory elements than other introns in *Drosophila melanogaster* (Duret 2001). Our dataset does not agree with this result, however, since we find that first introns are shorter than non-first introns, although short first introns have a lower divergence than non-first short introns. This is probably due to our only studying a limited set of genes, because when using the whole genome, first introns are significantly longer than non-first introns.

2.4.2 Intergenic sequences

As expected, the putative UTRs (edges of intergenic sequences that correspond to the average UTR lengths in *D. melanogaster*) have smaller divergence levels than the rest of the intergenic sequences, so they are under apparently more constraint than other non-coding sequence. Several studies have found that UTRs in *D. melanogaster* and *D. simulans* are under greater selective constraint than fourfold-synonymous sites (Halligan et al. 2004; Andolfatto 2005; Halligan and Keightley 2006; Haddrill et al. 2008a). As Casillas et al. (2007) point out, the indiscriminate use of any type of non-coding sequence as an unconstrained neutrally evolving standard is dangerous, given this evidence for differences in constraint levels among different classes. It may be better to use fourfold-synonymous sites or non-coding sequences that are thought to be under weak constraints, such as short introns.

2.4.3 X-linked versus autosomal loci

Singh et al. (2005a) found a strong negative correlation between codon bias and recombination rate on the *D. melanogaster* X chromosome, as opposed to the weak positive correlation found on the autosomes, suggesting that studies of coding sequence evolution should consider the X chromosome and autosomes separately. In addition, higher codon usage bias for X-linked loci has been previously found in an analysis of 9800 coding sequences from *D. pseudoobscura* that had orthologues in *D. melanogaster* (Singh et al. 2005b) and more recently in an analysis of 6698 genes using the 12 *Drosophila* genomes (Singh et al. 2008). These authors suggested that this pattern of higher codon usage on the X chromosome may be explained by one or both of the following two causes: the effective population size for females may be much higher than the effective size for males, or the selection on translational efficiency may be stronger on the X due to the hemizygosity of X-linked genes in males.

Given the fact that preferred codons in *Drosophila* end in G or C, a higher GC content, especially in the third position of codons, is thus expected in X-linked loci. In our set of genes, we also found that X-linked loci had a higher codon usage bias than autosomes, as well as a higher GC content, than autosomal loci. There is no difference in

the level of divergence between X-linked and autosomal loci at synonymous sites however.

In non-coding sequences, the only significant difference between X-linked and autosomal loci is a higher GC content in X-linked introns than in autosomal introns in long genes. This can be explained if there is stronger selection or biased gene conversion in X-linked genes affecting GC content at non-coding sites (Singh et al. 2005a), and the lack of significance in short genes might be due to a smaller sample size in this category.

2.4.4 Coding sequence patterns

One of the more interesting patterns that we recover in this study is a significant negative correlation between codon usage and non-synonymous substitution rates, even when the effects of gene expression levels, gene length, and synonymous divergence are taken into account. Lower codon usage in fast-evolving genes can be explained in several ways. Hill-Robertson interference (Hill and Robertson 1966) from strongly selected sites on the effects of weak selection for optimal codon usage at linked sites is frequently invoked (Betancourt and Presgraves 2002; Andolfatto 2007). This effect might be magnified in *D. miranda* by the evolution towards a lower overall codon usage bias (Bartolomé and Charlesworth 2006; Bachtrog 2007), caused by a lower effective population size than for *D. pseudoobscura*. This is because selection for codon usage is becoming weaker in *D. miranda* and thus strong positive selection will override the effects of selection for codon usage all the more. Bierne and Eyre-Walker (2006) claimed that Hill-Robertson effects do not greatly affect codon bias, and they argue that the only likely alternative is that the strength of selection acting upon synonymous mutations is correlated with that acting upon non-synonymous mutations, presumably because of selection on translational accuracy. Genes that are under greater selective constraint will evolve slowly and need to be accurately translated. Relaxed selective constraint on fast-evolving genes would then lead to a lower selection for codon usage.

Andolfatto (2007) showed that the negative correlation between codon usage and synonymous divergence between *D. melanogaster* and *D. simulans*, usually interpreted as being caused by lower divergence for more highly constrained synonymous sites,

disappears if the data are corrected for the correlation between nonsynonymous and synonymous divergence, supporting the interference hypothesis. He also showed that the magnitude of the effect is in agreement with what is expected from the observed rate of substitution of positively selected amino-acid mutations between *D. melanogaster* and *D. simulans*. Our results are in general agreement with this interpretation. Two recent studies of polymorphism and divergence in *D. miranda* and *D. pseudoobscura* suggest that a significant fraction of nonsynonymous divergence has been driven by positive selection (Bachtrog 2008; Haddrill et al. under review), as is required on this hypothesis. Drummond and Wilke (2008) found a negative correlation between F_{op} and d_S in *Drosophila melanogaster*, which they explain by selection against misfolding of proteins. This relationship is the only one in Drummond and Wilke (2008) with which our results do not agree, and the lack of a significant correlation between F_{op} and d_S could be due to the limited number of genes in our dataset. However, the Drummond and Wilke (2008) hypothesis predicts that gene expression differences drive all of the patterns that they find. Therefore, finding a negative correlation between F_{op} and d_N even after correcting for gene expression seems to suggest that either their hypothesis does not explain all aspects of the data, and that hitchhiking effects are involved, or that the gene expression dataset that we used does not capture all relevant features.

3 Polymorphism data shows that X-linked first introns and short introns in *Drosophila americana* are selectively constrained.

Contributing authors:

- I collected and analysed the data.
- A. Betancourt advised on the laboratory work and helped with analyses.
- B. Charlesworth advised on the project.

3.1 Introduction

A growing number of studies investigating evolutionary patterns of non-coding DNA sequences are showing evidence for natural selection acting on these sequences, which previously were assumed to be neutrally evolving with no selective constraint (Parsch 2003; Bergman and Kreitman 2001; Haddrill et al. 2005; Andolfatto 2005; Halligan and Keightley 2006; Casillas et al. 2007; Haddrill et al. 2008a). The functions of these sequences remain largely unknown but it has been suggested that they may have a role in gene regulation (Arnone and Davidson 1997; Hardison 2000; Parsch 2004), pre-mRNA secondary structure (Kirby et al. 1995; Leicht et al. 1995; Chen and Stephan 2003, Rogic et al. 2008) or RNA editing (Reenan 2005) and there is increasing experimental evidence for these (Birney et al. 2007). More recently, it has also been shown that natural selection on coding and non-coding DNA sequences is related to nucleosome organization (Warnecke et al. 2008; Babbitt and Kim 2008; Kaplan et al. 2009).

Intron and Untranslated Transcribed Regions (UTR) polymorphism datasets have been used to discriminate the types of selection acting on non-coding sequences in *Drosophila* (Andolfatto 2005; Begun et al. 2007; Casillas et al. 2007; Haddrill et al. 2008a). Purifying selection and even positive selection have now been established to affect non-coding sequences. Some patterns of non-coding sequence evolution, such as a negative correlation between intron divergence and intron length, have been shown to

hold for several *Drosophila* species (Haddrill et al. 2005; Bachtrog 2006; Marion de Procé et al. 2009). Another observation is that first introns generally have a higher frequency of conserved regulatory elements in *D. melanogaster* (Duret 2001) and in mammals (Majewski and Ott 2002), although mammalian and *Drosophila* introns may be evolving differently. First introns are also longer than other introns in *D. melanogaster* (Duret 2001; Marais et al. 2005; Bradnam and Korf 2008). The relationship between intron divergence and intron length may thus be affected by the positions of introns; although Haddrill et al. (2005) found that mean divergence did not differ between first and non-first introns within short- and long-intron size categories.

Drosophila americana and *Drosophila virilis* belong to the *virilis* group and are closely related, with a mean silent site divergence of 10.9% (Maside and Charlesworth 2007). *D. americana* has been a model for population genetic and evolutionary studies for several decades (Patterson and Stone 1952; Throckmorton 1982). In *D. americana*, some populations have the X chromosome and the 4^{th} chromosome fused, and an inversion has occurred in the same region. This makes it a particularly good model for early neo-sex chromosome evolution studies (McAllister and Charlesworth 1999; McAllister 2002; Vieira et al. 2006). This species has a well-defined ecology, independent of human activity (Throckmorton 1982), and might thus have a relatively stable demographic history, which allows the detection of natural selection without the interference of potential signals of demographic events. The *D. virilis* genome sequence is one of the 12 sequenced *Drosophila* genomes (Clark et al. 2007), allowing primer design, annotation of introns and calculation of divergence. For these reasons, it is good material for evolutionary genetic studies.

This work presents an analysis of 32 X-linked introns sampled from 14 lines of *Drosophila americana*. We use polymorphism and divergence analyses that allow tests for positive and negative selection, in order to assess whether patterns found in *D. melanogaster* and *D. simulans* hold for *D. americana*.

3.2 Materials and methods

3.2.1 Biological material

We used 14 *Drosophila americana* isofemale lines from the HI99 population from the South bank of Missouri River, near Howell Island Conservation Area, west of St. Louis, Missouri, with latitude 38° 39.7' N and longitude 90° 40.7' W (Fig. 3.1) (http://www.biology.uiowa.edu/mcallister/HI.html), kindly provided by Bryant McAllister. For this population, about 84.6% of the lines have the X-4 fusion, as estimated from 39 chromosomes analysed by Bryant McAllister (Vieira et al. 2001; McAllister 2002; McAllister and Evans 2006). All flies were maintained on standard banana medium (3.75L distilled water, 40g agar, 130g powdered malt, 115g yeast, 95mL syrup, 8 bananas, 18mL propionic acid, 30mL Tegosept) at 19°C.

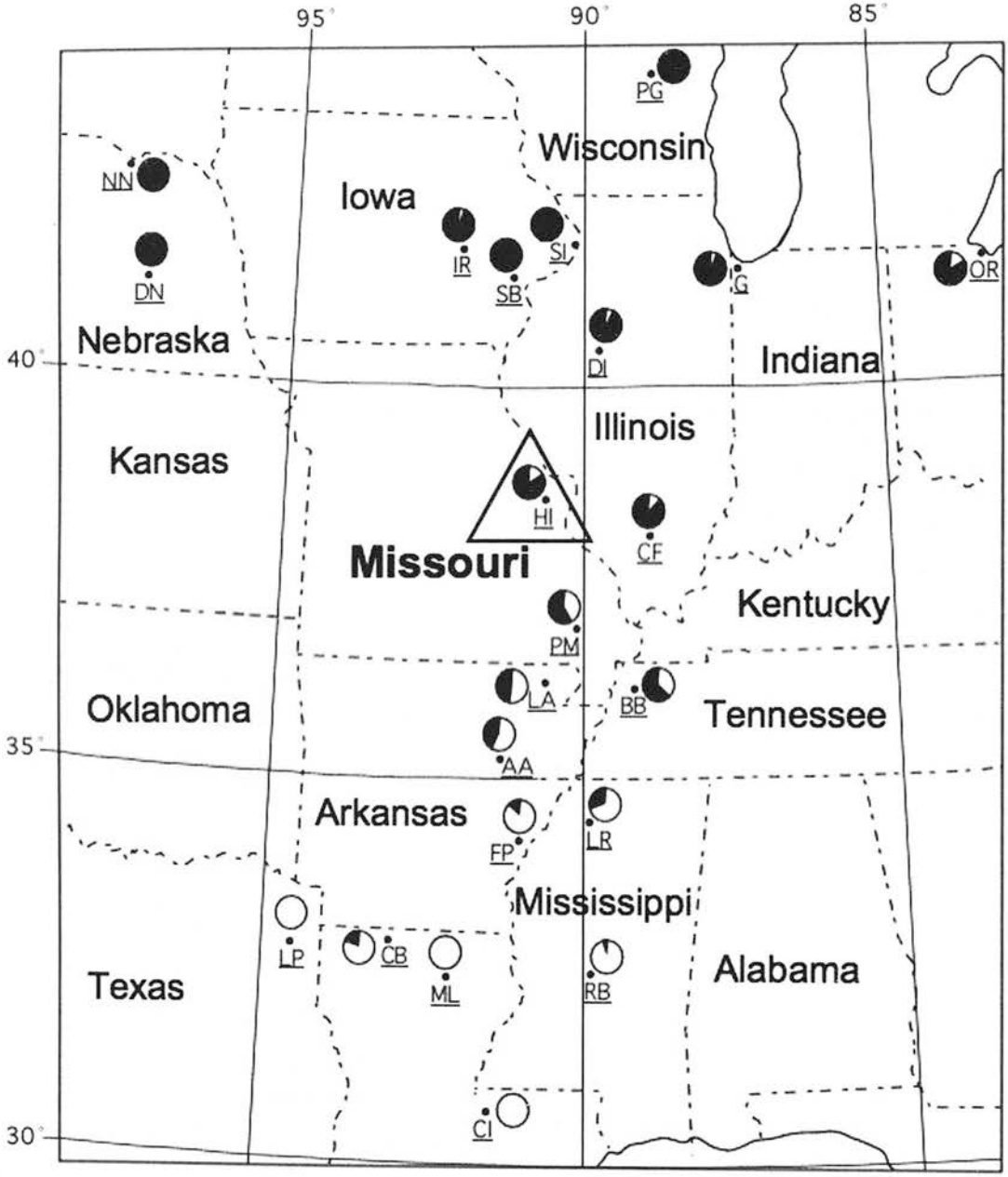

Figure 3.1: Map of the United States with the locations of *D. americana* populations. HI99 is located in Missouri, highlighted by the triangle symbol. For each locality, the frequency of the X-4 fusion chromosome in the sample is indicated as the portion of the circle that is filled and the frequency of unfused X chromosomes is indicated by the unfilled portion (http://www.biology.uiowa.edu/mcallister/HI.html) (Vieira et al. 2001; McAllister 2002; McAllister and Evans 2006).

3.2.2 Primer design

Introns located on the X chromosome were sequenced. We used primers from Maside et al. (2004) to obtain sequence from five short introns and a long intron (*csw*, *Pros28.1* and *Yp1*), and designed additional primers for long introns (about 500 bp). Because genes located near the fusion region or in inversions may have reduced variability that would interfere with signals of natural selection, regions affected by the *X/4* fusion or known segregating inversions (*Xa*, *Xb* and *Xc*), were excluded from this study. We identified these regions by combining information from Warters (1944), Hsu (1952), Patterson and Stone (1952) and FlyBase (http://flybase.org/maps/chromosomes/maps.html). The scaffolds remaining to design primers from the *D. virilis* sequence were the end of scaffold 13042 (coordinates 3,700,000 to 4,991,987) and almost all of scaffold 12928 (coordinates: 50,000 to 5,900,000) (Fig. 3.2). To determine the approximate location of the genes (Fig. 3.2), we used the map of the *D. virilis* X (Vieira et al. 2006) and searched for landmarks on the *D. virilis* genome through the DroSpeGe website (http://insects.eugenes.org/DroSpeGe/). This method assumes that there have not been many small-scale gene rearrangements after the divergence of *D. virilis* and *D. americana*.

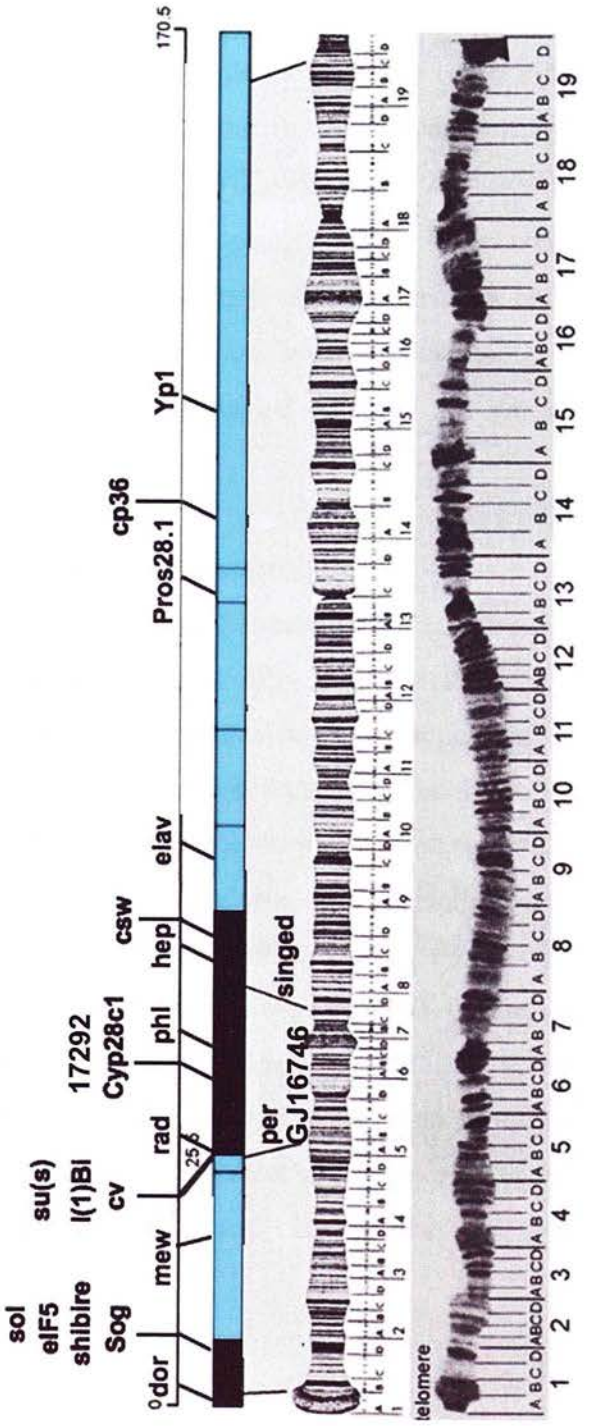

Figure 3.2: Map of the *D. virilis* X chromosome with locations of genes used in this study. Black regions correspond to the regions located away from known inversions. The map was obtained from the Chromosome Maps tool in Flybase (http://flybase.org/maps/chromosomes/maps.html).

3.2.3 Extraction and sequencing

We extracted genomic DNA from single males of all 14 HI99 lines using the Puregene (QIAGEN, West Sussex, UK) DNA purification kit and amplified fragments by PCR (conditions: 1.5-2mM $MgCl_2$; 94°C for 2'; 94°C for 30", 52-58°C for 30", 72°C for 1-2' (35 cycles); 72°C for 10'; 4°C). PCR products were then cleaned up using ExoSap-IT (USB corporation), which removes single-stranded primers and remaining nucleotides. Fragments were directly sequenced on both strands using the Big Dye (Applied Biosystems, Foster City, CA) sequencing kit and run on an ABI 3730 capillary sequencer.

3.2.4 Sequence alignment and processing

Sequence trace files were edited using Sequencher 4.7 (Gene Codes, Ann Arbor, MI) and aligned using MCALIGN2 (Wang et al., 2006). In long introns, we observed size polymorphisms due to insertions-deletions (indels) among lines. In order to exclude sites involved in splicing processes, we removed the first 7 bp at the 5' end and the last 7 bp at the 3' end of each intron, as these sites show the most constraint in divergence studies (Halligan and Keightley 2006; P. Haddrill and D. Halligan, pers. comm.). *D. virilis* sequences were obtained either from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) or from DroSpeGe (http://insects.eugenes.org/DroSpeGe/). The resulting dataset contains sequences for 32 introns sampled from 18 loci, including 12 short introns and 20 long introns. For 13 loci, there are pairs of first and second introns in the dataset. We also obtained the coding sequences for a subset of 15 genes, and retrieved 5 additional X-linked coding sequences from Maside and Charlesworth (2007) to compare synonymous sites and introns.

3.2.5 Data analysis

We used DnaSP 4.10 (Rozas et al. 2003) to obtain silent nucleotide diversity, Tajima's D, Fu and Li's D, the D/D_{min} ratio (Schaeffer 2002) and divergence from *Drosophila virilis*. We used a Jukes-Cantor (1969) correction to account for multiple hits and reverse mutations in calculations of diversity and divergence. The D/D_{min} ratio allows the comparison of Tajima's D between loci with different numbers of segregating sites. For these analyses, sites overlapping alignment gaps were excluded. In order to account for factors that can vary between genes and influence divergence and polymorphism patterns, we used paired t-tests to compare first and non-first introns of the same gene. We used Mantel-Haenszel (Mantel and Haenszel 1959) tests to compare ratios of polymorphism to divergence between different categories of sites (short vs. long introns, first vs. non-first introns, and all intron categories vs. synonymous sites) in the same gene (Appendix 3.1). This test compares the same number of two-by-two contingency tables of segregating sites and fixed differences for each category of sites as the number of loci used, and determines whether there is a consistent heterogeneity between the two categories of sites considered.

For each gene, we tested the difference in log likelihood between the neutral model and a model where this one gene was under selection using the MLHKA test (Wright and Charlesworth 2004) and we corrected the p-value with a Bonferroni correction (Sokal and Rohlf 1995).

3.3 Results

3.3.1 General

Intron characteristics, diversity and divergence indices, Tajima's D, Fu and Li's D and D/D_{min} ratio are presented in Table 3.1. The majority of the values for Tajima's D, Fu and Li's D and D/D_{min} were negative, although intron 1 in *dor*, introns 1 and 2 in *csw* and intron 4 in *sog* showed positive values for D_{FL}. The mean values for π_{JC} (2.17%) and θ_W

(2.32%) are slightly higher than those in Maside and Charlesworth (2007) (1.59% and 1.61% respectively).

Table 3.1: Characteristics of introns used in this study, ordered by their position on the *D. virilis* X chromosome. *L* is the average intron length in *D. americana*, *S* is the number of segregating sites, *D* is the number of fixed differences.

Gene	Intron	*N*	*L*	%GC	*S*	π_{JC} (%)	θ_W (%)	*D*	K_{JC} (%)	D_{FL}	D_T	D/D_{min}
dor	2	14	491.7	32.8	7	0.38	0.49	58	13.77	-0.81	-0.67	-0.32
dor	1	14	64	30	1	0.54	0.62	5	12.74	0.68	-0.34	-0.29
sol	4	14	1003	37.7	63	3.45	0.04	26	10.08	-0.54	-0.39	-0.75
sol	2	12	59	38.5	2	1.48	1.44	1	4.09	-0.48	-0.05	-0.03
sol	1	13	680.31	38.2	14	0.43	0.62	22	3.83	-1.86	-1.60	-0.71
eIF5	2	13	582.3	34.9	33	2.03	2.37	33	9.81	-0.86	-0.68	-0.35
eIF5	1	12	142.6	34.2	13	3.12	3.28	14	20.26	-0.01	-0.49	-0.69
shibire	6	14	752.3	39.9	67	3.30	3.61	29	10.36	-1.10	-0.80	-0.79
sog	4	14	574.9	36.6	61	4.27	3.84	45	14.42	0.12	0.20	-0.01
mew	5	14	729.8	34.5	60	3.70	3.78	32	10.41	-0.90	-0.51	-0.34
l(1)1Bi	2	14	64	33.5	2	0.99	1.21	4	8.73	-0.54	-0.53	-0.35
l(1)1Bi	1	14	829.9	39.2	64	2.88	3.16	39	8.97	-0.43	-0.55	-0.18
cv	1	14	322.8	38.3	13	1.01	1.14	11	4.46	-1.00	-1.30	-0.48
cv	2	14	57	19.4	5	2.92	3.57	4	14.43	-0.35	-0.78	-0.40
rad	1	14	232.29	52.5	17	2.42	1.62	12	7.98	-0.42	-0.56	-0.06
rad	2	13	501.43	35.3	31	2.56	2.53	17	9.95	-1.13	-0.37	-1.06
GJ16746	1	13	411.6	40.6	28	1.95	2.47	34	12.57	-1.49	-1.20	-0.51
GJ16746	2	14	63.1	38.7	5	2.27	3.15	7	20.10	-0.81	-1.54	-0.56
17292	1	14	773.6	39.8	20	0.74	0.86	28	4.95	-0.61	-0.71	-0.31
17292	2	13	62	47.9	2	0.65	1.32	0	0.32	-1.96	-1.47	-0.97
Cyp28c1	2	14	461.4	35.4	48	2.90	3.58	22	9.89	-0.85	-1.00	-0.60
Cyp28c1	1	14	523.7	34.6	77	4.41	5.00	35	15.26	-2.04	-0.99	-0.89
phl	4	14	917.29	42.3	65	3.23	3.55	38	14.81	-0.75	-0.48	-1.17
si	6	14	630.5	40.9	49	2.50	3.19	44	13.19	-1.32	-1.24	-0.51
hep	2	14	61	21.8	5	2.51	3.15	7	18.23	-1.17	-0.78	-0.40
hep	1	14	581.8	40.4	26	1.28	1.48	25	6.58	-0.71	-0.72	-0.36
csw	1	14	74.8	38.3	3	1.62	1.85	2	4.70	0.68	-1.22	-0.56
csw	2	14	86.9	42.9	10	5.36	5.13	4	9.73	0.57	0.68	0.31
Pros28.1	1	14	77.8	31.6	1	0.23	0.49	0	0.11	-1.48	-1.16	-0.97
Pros28.1	2	14	67	31.7	2	0.78	1.17	3	6.32	-0.54	-0.96	-0.63
Yp1	1	14	68	35.2	3	1.03	1.72	1	3.80	-1.21	-1.28	-0.74
Yp1	2	14	79.1	30.3	6	2.43	2.86	7	14.50	-0.08	-0.66	-0.33
Average (SE) / Total				36.5 (1.1)	803	2.17 (0.23)	2.32 (0.22)	609	9.98 (0.92)	-0.73 (0.12)	-0.75 (0.09)	-0.50 (0.06)

Tables 3.2 and 3.3 present diversity and divergence indices, and Tajima's *D* for synonymous and non-synonymous sites. Our values for π_{JC} (1.96%) and θ_W (1.77%) for synonymous sites are very close to those from Maside and Charlesworth (2007) (1.96% and 2.10% respectively). For non-synonymous sites, our values for π_{JC} (0.09%) and θ_W

(0.11%) are slightly higher than theirs (0.04% and 0.05% respectively), but still fairly close.

Table 3.2: Characteristics of the synonymous sites on X-linked coding sequences used in this study, ordered by their position on the *D. virilis* X chromosome.

Gene	N	L	S	π_{JC} (%)	θ_W (%)	D_T	D	$K_{S(JC)}$ (%)
dor	14	185.6	4	0.87	0.68	0.88	22	13.50
sol	13	209.7	9	1.03	1.34	-1.01	17	9.82
eIF5	12	204.4	22	2.57	3.56	-1.29	15	10.71
sog	14	16.2	0	0	0	NA	1	0
su(s)	5	275.1	13	1.95	NA	-0.75	33	14.52
l(1)1Bi	14	129.1	13	2.76	3.17	-0.59	11	12.89
cv	14	61.8	5	2.20	2.55	-0.51	5	12.83
per	5	236.2	9	1.70	1.81	-0.53	31	15.36
rad	14	28.1	0	0	0	NA	3	11.50
GJ16746	14	93.5	6	1.32	2.02	-1.27	3	5.96
17292	14	48.3	0	0	0	NA	4	8.77
Cyp28c1	14	105.9	16	3.62	4.75	-1.06	6	8.19
si	14	11.3	0	0	0	NA	0	0
hep	14	107.8	6	1.36	1.75	-0.83	9	10.89
csw	14	219.8	18	1.85	2.58	-1.20	18	9.66
elav	50	227.3	27	1.58	2.64	-1.35	27	14.24
Pros28.1	14	160.1	7	11.14	1.36	-0.71	12	8.60
Cp36	5	224.7	3	0.63	0.64	-0.18	12	6.22
Yp1	13	159.8	15	2.63	3.03	-0.61	16	13.29
Average (SE) / Total			173	1.96 (0.56)	1.77 (0.32)	-0.73 (0.13)	245	9.84 (1.01)

Table 3.3: Characteristics of the non-synonymous sites on X-linked coding sequences used in this study, ordered by their position on the *D. virilis* X chromosome.

Gene	N	*L*	*S*	π_{JC} (%)	θ_W (%)	D_T	*D*	$K_{A(JC)}$ (%)
dor	14	555.4	2	0.09	0.11	-0.53	4	0.78
sol	13	648.3	6	0.20	0.34	-1.38	11	1.83
eIF5	12	695. 6	1	0.02	0.05	-1.14	0	0.01
sog	14	46.8	0	0	0	NA	0	2.17
su(s)	5	838	NA	0.59	NA	-0.30	22	3.38
l(1)1Bi	14	410.9	4	0.17	0.31	-1.48	3	0.82
cv	14	238.2	0	0	0	NA	1	0.42
per	5	771.8	3	0.16	0.19	-1.05	13	1.94
rad	14	85.9	0	0	0	NA	0	0
GJ16746	14	323.5	3	0.13	0.29	-1.67	1	0.38
17292	14	137. 7	0	0	0	NA	0	0
Cyp28c1	14	320.1	3	0.13	0.30	-1.67	3	1.01
si	14	48. 7	0	0	0	NA	0	0
hep	14	360.2	0	0	0	NA	0	0
csw	14	692.2	3	0.11	0.14	-0.57	1	0.21
elav	50	705.7	0	0	0	NA	0	0
Pros28.1	14	499.9	0	0	0	NA	5	1.01
Cp36	5	627.3	1	0.10	0.08	1.23	2	0.38
Yp1	13	557.2	3	0.08	0.17	-1.65	2	0.40
Average (SE)			29	0.09 (0.03)	0.11 (0.03)	-0.93 (0.20)	68	0.78 (0.22)

3.3.2 Divergence

Overall, there was little evidence for differences in divergence between different classes of silent sites. There is no significant difference in divergence between any intron class and synonymous sites (Wilcoxon rank sum tests, $p>0.13$).

Intron divergence (after Jukes-Cantor correction) is not significantly correlated with intron length (Fig. 3.1.A) (Spearman's rank correlation rho r_S = 0.0733, 95% bootstrap C.I.: [-0.311; 0.425]). After correcting for GC content, the correlation between intron divergence and intron length is slightly higher but still not significant (Spearman's rank correlation rho r_S = 0.1885, 95% bootstrap C.I.: [-0.198; 0.527]). As opposed to previous results in different *Drosophila* species (Haddrill et al. 2005; Bachtrog and Andolfatto 2006; Marion de Procé et al. 2009), we find that long introns have a larger mean divergence (K_i = 0.106) than short introns (K_{JC} = 0.090), although this difference is not significant (Wilcoxon test; W= 97, p = 0.39).

Figure 3.3.A suggests a tendency for first introns to have a lower divergence than other introns. We therefore tested the difference in divergence between first and non-first introns. First introns have a lower divergence than non-first introns (first: 0.082, non-first: 0.112) but the difference is not significant (Wilcoxon test, $W=76$, $p = 0.07$). Using pairs of introns, first introns have a lower divergence than second introns in the same gene (first: 0.082, second: 0.108) but the difference is not significant (paired t-test, $t = -1.38$, $p = 0.19$).

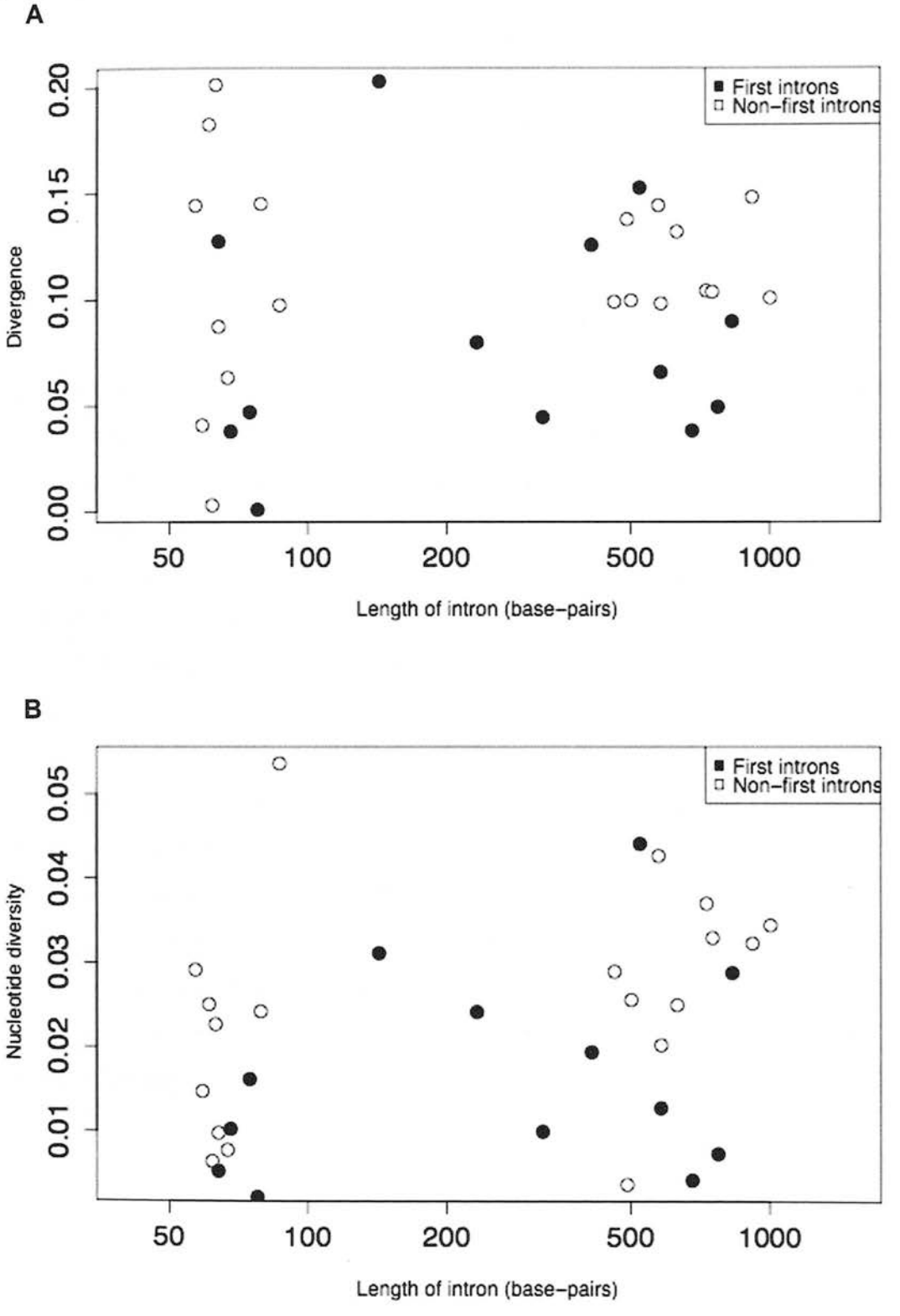

Figure 3.3: **A**. Plot of intron divergence (with a Jukes Cantor correction) against intron length (on a log scale). **B**. Plot of intron nucleotide diversity (π_{JC}) against intron length (on a log scale). Dark circles represent first introns and open circles represent non-first introns.

3.3.3 Polymorphism

There is no significant difference in polymorphism between any intron class and synonymous sites (Wilcoxon rank sum tests, $p>0.13$), although the mean value for π_{JC} for all introns (2.17%) is higher than that for synonymous sites (1.96%).

When we look at intron length and polymorphism (after Jukes-Cantor correction) (Fig. 3.3.B), we find a positive correlation, which is close to significance (Spearman's rank correlation rho = 0.3314, 95% bootstrap C.I.: [-0.0003; 0.6146]). The difference in mean polymorphism (after Jukes-Cantor correction) between short and long introns is significant (Wilcoxon rank sum test, W=60, p = 0.019), with long introns having a higher polymorphism (long: 0.026, short: 0.015). These results are consistent with the relationship between intron length and divergence, but again in the opposite direction to other *Drosophila* studies.

The same test for polymorphism level (after Jukes-Cantor correction) between first and non-first introns is also not significant (W = 75, p-value = 0.065) but in the same direction as the divergence result (first: 0.017, non-first: 0.025). First introns have a lower polymorphism than second introns in the same gene (first: 0.017, second: 0.021) but the difference is not significant (paired t-test, t=-1.03, p=0.32).

We investigated the relationship between intron divergence (K_{JC}) and intron polymorphism (π_{JC}) (Fig. 3.4). There is a significant positive correlation between K_{JC} and π_{JC} (Spearman r_S=0.58, p=0.0006), consistent with the neutral theory prediction (Kimura 1983). However, it seems that the data is well spread, suggesting that individual introns evolve differently.

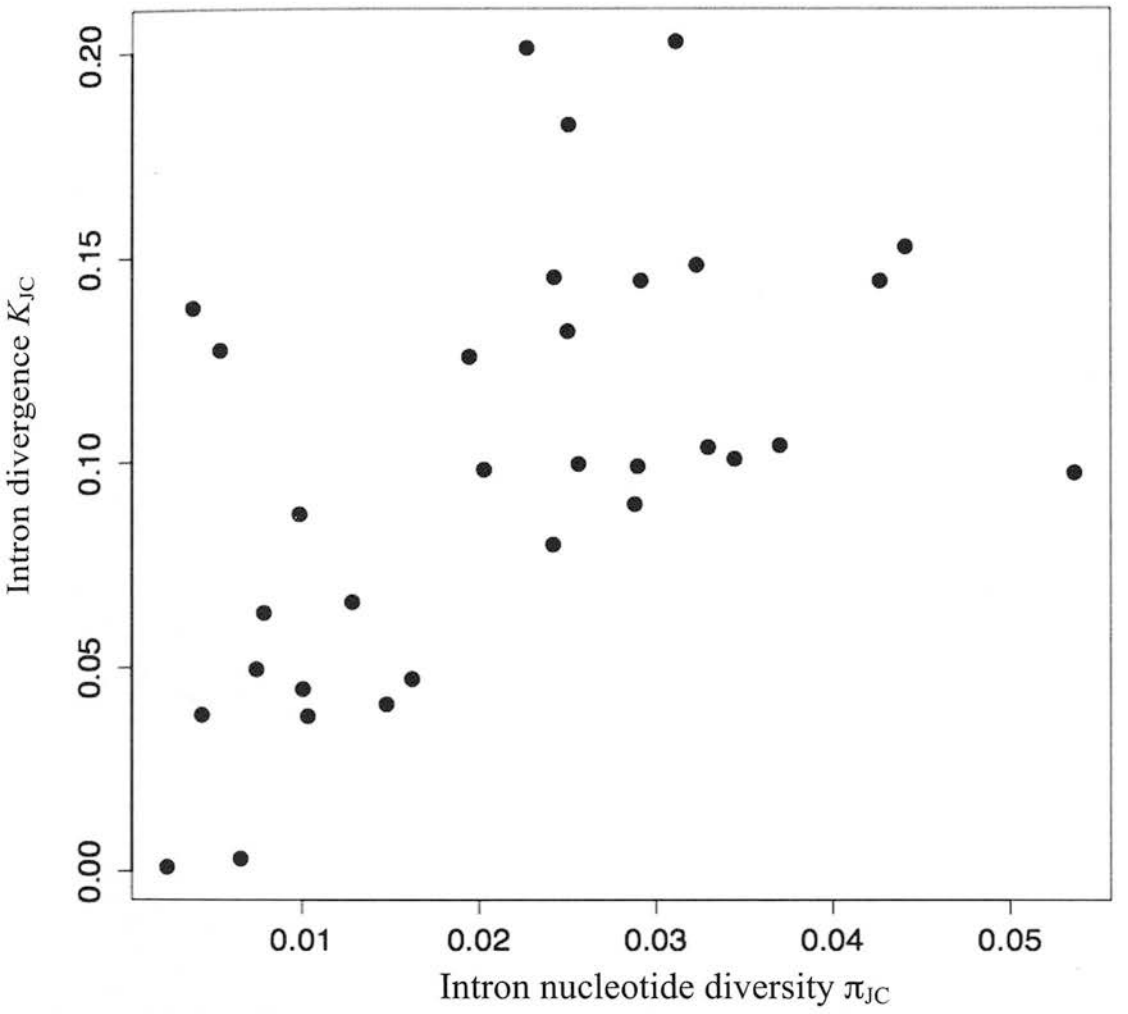

Figure 3.4: Plot of intron divergence K_{JC} against intron nucleotide diversity π_{JC}.

3.3.4 Tajima's *D* and D/D_{min} ratio

There is no significant difference in Tajima's *D* or D/D_{min} ratio between any category of sequences (Fig. 3.5 and 3.6). There is no significant difference in Tajima's *D* between any intron class and synonymous sites (Wilcoxon rank sum tests, $p>0.365$). Non-synonymous sites have a lower Tajima's *D* than introns and synonymous sites, but the difference is not statistically significant (Wilcoxon tests, $p>0.17$). Similarly, there is no significant difference in D/D_{min} ratio between any intron class and synonymous sites (Wilcoxon rank sum tests, $p>0.10$). The D/D_{min} ratio for all introns pooled is lower than

the D/D_{min} ratio for synonymous sites but this difference is not significant (Wilcoxon rank sum test: p=0.0783).

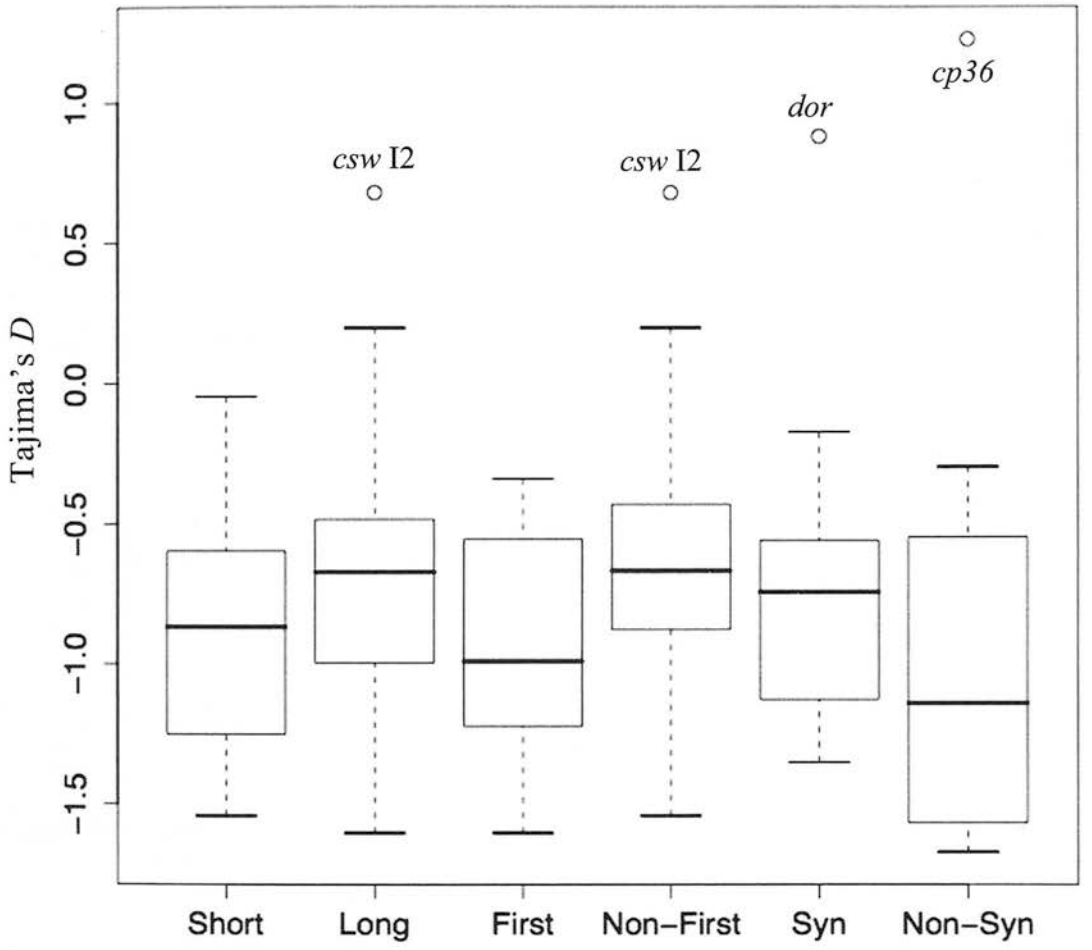

Figure 3.5: Boxplot of Tajima's *D* for different classes of introns and coding sites. White circles represent outliers, ie. values beyond the third quartile (Q_3) plus 1½ times the interquartile range ($IQR=Q_3-Q_1$).

There is no significant difference in Tajima's *D* or D/D_{min} ratio between short and long introns (Wilcoxon rank sum tests, p>0.26) or between first and non-first introns (Wilcoxon rank sum tests, p >0.13).

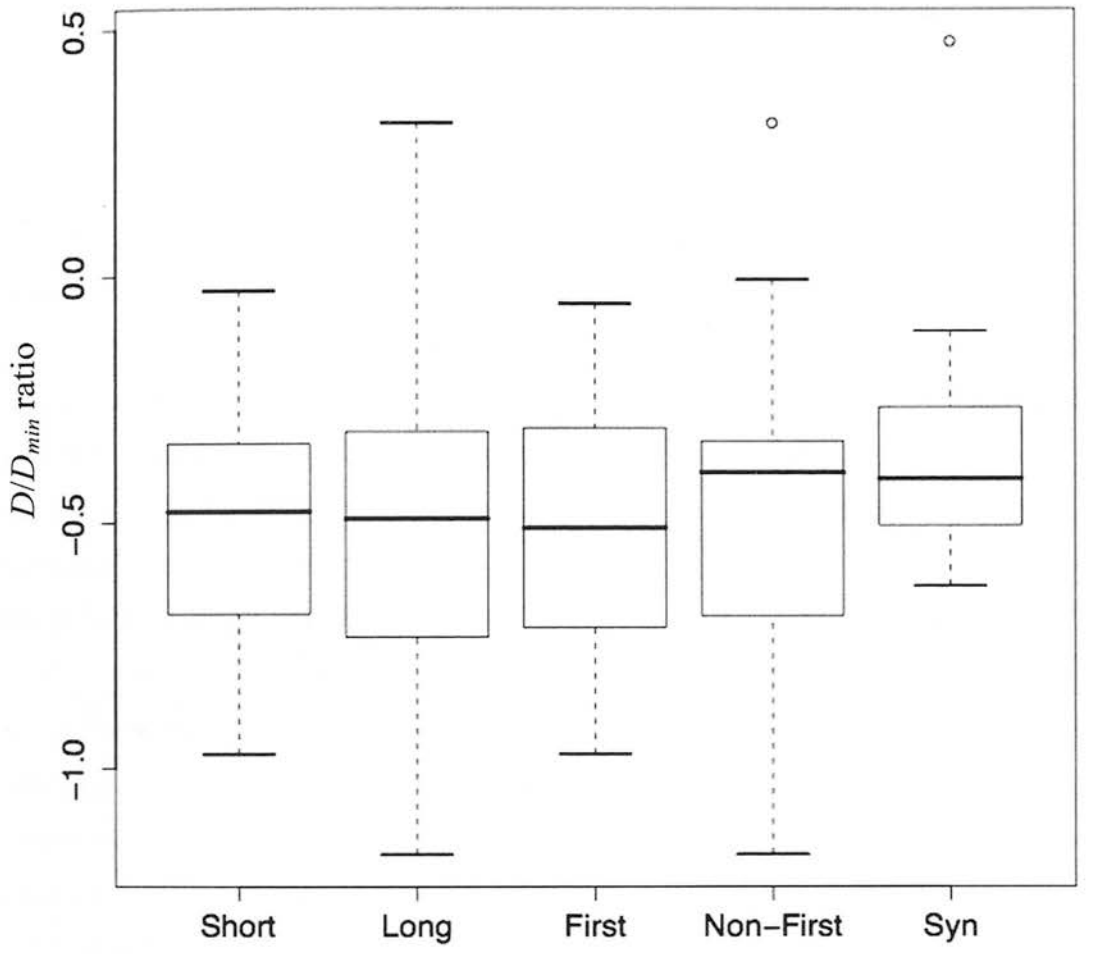

Figure 3.6: Boxplot of the D/D_{min} ratio for different classes of introns and synonymous sites. White circles represent outliers, ie. values beyond the third quartile (Q_3) plus 1½ times the interquartile range ($IQR=Q_3-Q_1$).

3.3.5 Tests for natural selection

We first tested whether any single intron was evolving differently from all other introns using the MLHKA test (Wright and Charlesworth 2004); no model assuming that one intron was selected gave a significantly higher likelihood than the model where all introns are considered neutrally evolving using likelihood ratio tests.

We tested the difference in polymorphism to divergence ratio for every class of introns against synonymous sites using McDonald-Kreitman tests. We first tried to detect any purifying selection, so we included all polymorphisms for the McDonald-Kreitman tests. Then we tested for positive selection, excluding singletons. The rationale for this is

that purifying selection will lead to an excess of rare variants, so we need to include these to detect purifying selection. Positive selection, on the other hand, should increase divergence relative to polymorphism levels. Therefore, removing singletons, which only contribute to polymorphism if they are negatively selected, gives more power to the detection of positive selection. The mean ratio of π_{JC} over K_{JC} (r_{PD}) for introns is the same in this dataset (0.20) as in the dataset used in Haddrill et al. (2008a) (0.21).

None of the Mantel-Haenszel tests against synonymous sites were significant ($p>0.11$) (Table 3.4). This suggests that all categories of introns are subject to similar constraint to synonymous sites. However, if we look at the r_{PD} ratio, the value for short introns is lower than that for synonymous sites, suggesting that short introns have lower polymorphism levels than synonymous sites for similar levels of divergence, even though this difference is not significant.

As our results showed that short introns seem to have lower polymorphism and lower divergence than long introns, we would expect short introns to be more constrained than long introns. We used a Mantel-Haenszel test to determine whether short and long introns in the same gene had consistently different ratios of polymorphism to divergence, and there was no significant heterogeneity between contingency tables ($p = 0.88$).

Since we observed a trend for lower divergence and polymorphism for first introns, we also used a Mantel-Haenszel test to determine whether first and second introns in the same gene had consistently different ratios of polymorphism to divergence. There was no significant heterogeneity between contingency tables ($p = 0.97$).

Table 3.4: Polymorphism and divergence in different classes of sites. r_{PD} is the ratio of polymorphism over divergence.

Sequence Class	Total No. Loci	Total No. bp	No. loci compared	Mean π[a] (SE)	Mean K_{JC}[b] (SE)	r_{PD}	$P^{all(c)}$	$P^{1(c,d)}$
Synonymous	19	2,704	–	1.96 (0.56)	9.84 (1.00)	0.20	–	–
Introns	32	9,690	15	2.17 (0.23)	9.98 (0.92)	0.22	0.129	0.138
Short introns	12	622	12	1.45 (0.26)	9.01 (1.97)	0.16	0.625	0.869
Long introns	20	9,068	17	2.60 (0.30)	10.56 (0.91)	0.25	0.123	0.114
First introns	13	4,288	13	1.67 (0.34)	8.17 (1.56)	0.20	0.417	0.731
Non-first introns	19	5,402	15	2.51 (0.29)	11.22 (1.07)	0.22	0.926	0.189
First / Second	–	–	13	–	–	–	0.951	0.944
Short / Long	–	–	8	–	–	–	0.876	0.593

[a] π_{JC} is the average Jukes-Cantor corrected pairwise divergence per nucleotide site between alleles (%).
[b] K_{JC} is the average Jukes-Cantor corrected divergence from *D. virilis* (%).
[c] P^{all}/P^{1} are probabilities from Mantel-Haenszel tests against synonymous sites including all polymorphisms/excluding singletons. The last two rows give the p-values for MH tests of first vs. second introns and short vs. long introns respectively.
[d] Excluding singletons did not alter the conclusions.

3.4 Discussion

Our results suggest that short introns are likely to be affected by purifying selection, as well as first introns to a lesser extent. This contrasts with previous studies that showed higher levels of constraint on long introns.

3.4.1 Divergence

In contrast to large scale studies in *D. melanogaster* and *D. simulans* (Halligan et al. 2004, Andolfatto 2005, Haddrill et al. 2005, Marais *et al* 2005, Halligan and Keightley 2006, Casillas *et al* 2007, Haddrill *et al* 2008a) and *D. miranda* (Marion de Procé et al. 2009), intron divergence is similar to divergence in synonymous sites in our dataset.

We do not find any significant relationship between intron divergence and intron length, contrary to the negative correlation between intron length and intron divergence

found in *D. melanogaster* and *D. simulans* (Haddrill et al. 2005; Halligan and Keightley 2006). If anything, the relationship is in the opposite direction, with long introns having a slightly higher divergence than short introns. The lower divergence observed in first introns when compared with non-first introns is not significant, as was also found by Haddrill et al. (2005).

3.4.2 Polymorphism

There is no significant difference in level of polymorphism between any intron class and synonymous sites. Introns in our dataset therefore have similar levels of polymorphism as synonymous sites. Given evidence for selection in synonymous sites (Maside and Charlesworth 2004 and Chapter 5), this lack of difference in polymorphism, as well as in divergence levels, implies selection or biased gene conversion on non-coding sites. This conclusion is confirmed by the analyses described in Chapter 5.

As for the divergence results, we find a positive correlation between intron length and polymorphism. This is not significant, but the mean level of polymorphism in long introns is significantly higher than in short introns. This correlation is in the opposite direction to results in previous *Drosophila* studies (Haddrill et al. 2005; Bachtrog and Andolfatto 2006).

The difference between polymorphism in first and non-first introns is significant and in the same direction as the divergence results; first introns have lower polymorphism than non-first introns. The difference is in the same direction in the paired intron dataset but is not significant. This suggests that purifying selection acts on first introns.

3.4.3 Tajima's D and D/D_{min} ratio

There is no significant difference in Tajima's D or D/D_{min} ratio between any category of sequences, suggesting that there is no difference in allele frequency spectrum between synonymous sites and all categories of introns investigated here, consistent with the conclusions discussed above. The average values of Tajima's D for all introns and synonymous sites are both negative and very similar (Tables 3.1 and 3.2), further suggesting that introns are subject to similar levels of purifying selection as synonymous sites.

3.4.4 Tests for natural selection

The HKA test results showed that no single intron was evolving differently from all other introns, suggesting that all introns are subject to similar evolutionary forces, whether they involve natural selection or not. However, Fig. 3.4 showed that introns have a lot of variation in polymorphism to divergence ratios; this difference could be due to different mutation rates between loci, or differences in the strength of weak purifying selection. It is possible that the HKA test does not have enough power to detect small differences in weak purifying selection between introns.

The Mantel-Haenszel test allows the comparison of polymorphism to divergence ratio in two classes of sites, one putatively neutrally evolving and one putatively selected. If these ratios are consistently different between the two classes of sites across all loci tested, it means that the putatively selected class of sites has undergone selection. We tested the difference in polymorphism to divergence ratio for every class of introns and synonymous sites. None of the Mantel-Haenszel tests against synonymous sites were significant, suggesting that all categories of introns are evolving similarly to synonymous sites. In Chapter 5, we show that synonymous sites are subject to selection for codon usage bias. This suggests that intron sites might also be subject to natural selection. However, if we look at the r_{PD} ratio, the value for short introns is lower than that for synonymous sites, suggesting that short introns have lower polymorphisms than synonymous sites for similar levels of divergence. Lower levels of polymorphism are expected when purifying selection on a locus removes the linked variation (Charlesworth

et al. 1993), suggesting that short introns could be subject to stronger purifying selection than synonymous sites. As sites in short introns are closer to coding sequence on average than sites in long introns, it is possible that this is an effect of purifying selection on neighbouring exons (see Chapter 5; Hill and Robertson 1966). However, this difference is not significant and might be due to the lack of power. Even though the results are not significant, the trends we observe for short introns are different from those in other *Drosophila* species, where longer introns showed lower divergence. This suggests that non-coding DNA in *Drosophila americana*, which is distantly related to *D. melanogaster*, might be evolving differently.

4 Insertions and deletions in an intron polymorphism dataset in *D. americana* and *D. simulans*.

Contributing authors:

- I collected the data and performed the analyses.
- K. Zeng performed the Maximum-Likelihood method.
- P. Haddrill provided the alignments for *D. simulans*.
- B. Charlesworth advised on the project.

4.1 Introduction

Insertions and deletion events are common in non-coding sequences, generally due to microsatellites, transposable elements or replication errors. Since they make up the majority of most genomes, variation in the amount of non-coding sequences is a major contributor to variation in the size of genomes. The relative amounts of deletions and insertions may influence non-coding DNA length and hence genome size. There is a weak positive relationship between intron size and genome size among *Drosophila* species (Moriyama et al. 1998; reviewed in Lynch 2007) and a significant correlation between transposable element-derived genomic DNA content and euchromatic genome size across *Drosophila* genomes (Clark et al. 2007), but no single class of non-coding DNA can explain alone the differences in genome size. Assuming non-coding sequences are mostly neutrally evolving with no selective constraint, then the relative mutation rates to deletions and insertions may be the primary influence on genome size.

In *Drosophila*, there appears to be a strong bias toward deletions. Petrov et al. (1996) observed a high rate of DNA loss in "dead-on-arrival" non-LTR retrotransposons, which evolve similarly to pseudogenes, in the *Drosophila virilis* group. Petrov and Hartl (1998) found that this mutational bias towards deletions is also present in the *Drosophila*

melanogaster group, which diverged from the *D. virilis* group 40 Mya, suggesting that this deletion bias is a general feature of *Drosophila*. A polymorphism deletion bias (PDB) has been further documented in *Drosophila melanogaster* in an intron polymorphism study (Comeron and Kreitman 2000; Ometto et al. 2006) and in mutation accumulation studies (Haag-Liautard et al. 2007; Keightley et al. 2009). These findings have led to a debate on whether directional selection for compact genome size could be responsible for the deletion bias (Charlesworth 1996; Petrov and Hartl 1997). However, there are two reasons to think that mutation, and not selection drives the deletion bias. First, Petrov and Hartl (2000) suggested that this genome-wide selection cannot solely explain the deletion bias, especially for small deletions, based on previous studies and theoretical considerations suggesting that natural selection for global genome size will have a negligible effect on small indels. Secondly, the deletion bias observed in mutation accumulation studies (Haag-Liautard et al. 2007; Keightley et al. 2009) definitely cannot be explained by natural selection.

For introns, in particular, several models have been proposed that involve putative selective pressures on insertions or deletions. Some invoke introns as potential modifiers of the recombination rate. Indeed, Comeron and Kreitman (2000) propose that insertions are selectively favoured as enhancers of recombination by increasing the physical distance and thus the probability of recombination between coding sequences, so this effect should be particularly strong in regions of low recombination. Other models focus on intron size, noting that introns with intermediate lengths are more efficiently spliced than either longer or shorter introns. According to Carvalho and Clark (1999), insertions are weakly deleterious and therefore more likely to be fixed in regions of low recombination. Ptak and Petrov (2002) inferred that indels affecting intron splicing or other functional constraint in introns, particularly long deletions, are subject to strong purifying selection and so are quickly eliminated, while the remaining indels are nearly neutral and can persist as polymorphisms. Parsch (2003) argues that the apparently deletion-biased mutation pressure should be balanced by favourable, compensatory insertions that restore optimal intron size.

Many studies have attempted to quantify the deletion bias. The deletion bias observed in the phylogenetic approach by Petrov and Hartl (1998) was approximately 8 deletions to 1 insertion. In an analysis of 31 genomic regions from various sources in *Drosophila melanogaster*, Comeron and Kreitman (2000) found an overall polymorphism deletion bias of 1.35 in introns and intergenic regions, and that this value was not significantly different for regions with different recombination rates. The difference in deletion bias between introns and pseudogenes may be accounted for by selective constraint on introns (Ptak and Petrov, 2002). Schaeffer (2002) studied the *Adh* region in *D. pseudoobscura* and found a polymorphism deletion bias of 1.89 for nonrepetitive indels. The polymorphism deletion bias (PDB) obtained by Ometto et al. (2005) for nonrepetitive indels was 2.00 for introns, close to the estimate by Schaeffer (2002). Parsch (2003) studied an intron polymorphism dataset in the *D. melanogaster* subgroup and found a deletion bias of 1.66.

Of course, the relative amount of DNA loss and gain is also determined by the size of the indels. Comeron and Kreitman (2000) and Parsch (2003) found that the vast majority of indels are shorter than 10bp. In a study of 22 intergenic sequences and 54 introns in *D. melanogaster*, Ometto et al. (2005) observed that insertions were smaller on average than deletions, which should accentuate the loss of DNA, but insertions also had higher mean frequencies than deletions, suggesting that they may be favoured to compensate for the DNA loss. Presgraves (2006) analysed an intron dataset for three closely related *Drosophila* species and found that small insertions were segregating at elevated frequencies compared with deletions, and they also had elevated probabilities of fixation.

In this Chapter, we have studied the sizes and frequencies of insertions and deletions in two intron polymorphism datasets, one for *D. americana* and one for *D. simulans*. *D. americana* is closely related to *D. virilis* (see Chapter 3), and *D. melanogaster* can be used as an outgroup for *D. simulans*. These species are thought to have been demographically stable for a long time (*D. americana*: Masidc and Charlesworth 2007; Madagascan *D. simulans*: Dean and Ballard 2004), which allows a better detection of potential signals of natural selection, although we find that *D.*

americana has been subject to a recent population expansion. Indeed, population size changes can affect the DNA sequences in similar ways to evolutionary forces, so that it is often difficult to distinguish them (Hahn et al. 2002).

4.2 Materials and methods

4.2.1 Fly species populations

For the *Drosophila americana* dataset, we used essentially the same data as were presented in Chapter 3. Direct sequencing gives high-quality DNA sequences and is particularly suited to insertion and deletion studies, unlike some next generation sequencing methods such as 454 or HeliScope (Shendure and Ji, 2008). There were 21 introns from 17 genes that included insertions and deletions, with 4 short and 17 long introns (using 80bp as a threshold) and 10 first and 11 non-first introns.

The *Drosophila simulans* dataset is described in Haddrill et al. (2008a). It included 20 individuals from a Madagascan population. There were 24 X-linked introns from 24 genes.

4.2.2 Analyses

Alignments were done and checked as described in Chapter 3. We removed 7 bp at the 5' and 7 bp at the 3' end of each intron to remove constraint from splicing sites. The first step was to determine "simple" insertions and deletions, meaning non-overlapping and non-contiguous events, corresponding to the non-repetitive class of indels used by Schaeffer (2002). This process removed a large number of insertions and deletions but it ensures that there are no ambiguous data or repetitive sequences that might be evolving differently. Figure 4.1 shows an example of a simple deletion and a complex indel event. We then polarised the polymorphisms using *D. virilis* as an outgroup for *D. americana*, and *D. melanogaster* as an outgroup for *D. simulans*. Hence, if we found two indel variants in a population, the variant found in the outgroup was the ancestral state. We calculated the size and frequency for each insertion and deletion.

To obtain maximum-likelihood estimates of the selection coefficient γ ($2N_e s$) for insertions and deletions, we used the method described in Zeng and Charlesworth (2009). We first estimated demographic parameters using a dataset of mutations at synonymous and non-coding sites. For *D. americana*, this dataset includes 20 X-linked coding sequences, in which we looked at preferred (P) and unpreferred (U) mutations, and 32 X-linked introns, in which we looked at GC and AT mutations. This method uses an approach similar to that of Cutter and Charlesworth (2006) and does not require polarising ancestral *vs* derived states. Instead, it only requires information on preferred and unpreferred states in the case of codon usage bias, and on GC and AT mutations in the case of non-coding DNA. The use of both fixed and polymorphic sites allows the distinction between mutational bias and patterns due to natural selection. To model changes in population size, the method allows for a sudden population size change t generations ago from N_a to N_b. The parameters estimated in this model are the selection coefficient γ (= $2N_b s$), the mutational bias κ, the nucleotide diversity θ (= $4N_b\mu$), the ratio of population size after expansion over the initial population size g (= N_a/N_b) and the time since population expansion τ (= t/N_a). The method assumes that AT to GC mutations are favoured (due to biased gene conversion) in the intron dataset and selection for U to P mutations in the coding sequence dataset due to translational selection.

The method then estimates γ for insertions and deletions conditional on the demographic parameters. This method allows the detection of natural selection in populations that have not been demographically stable in the recent past. Datasets with an excess of rare variants give large and negative γ estimates, whereas datasets with many indels at intermediate frequencies will yield large and positive γ estimates (Akashi 1999; Zeng and Charlesworth 2009).

```
HI9928    AACTAGTTTGTGAATATCGTTCCTTA-----TACTAGCTCACCGTTTGTTTGCT
HI9958    AACTAGTTTGTGAATATCGTTCCTTA-----GACTAGTTCACCGTTTGTTTGCT
HI9912    AACTAGTTTGTGAATATCGTTCCTTA-----TACTAGTTCACCGTTTGTTTGCT
HI9906    AACTAGTTTGTGAATATCGTTCCTTA-----GATTAGTTCACCGTTTG----CT
HI9904    AACTAGTTTGTGAATATCGTTCCTTA-----GACTAGTTCACCGTTTGTTTGCT
HI9946    ----------TGAATATCGTTCCTTA-----GACTAGTTCACCGTTTGTTTGCT
HI9914    AACTAGTTTGTGAATATCGTTCCTTA-----GACTAGTTCACCGTTTGTTTGCT
HI9924    AACTAGTTTGTGAATATCGTTCCTTA-----GACTAGTTCACCGTTTGTTTGCT
HI9934    AACTAGTTTGTGAATATCGTTCCTTA-----GACTAGTTCATCGTTTGTTTGCT
HI9918    AACTAGTTTGTGAATATCGTTCCTTA-----GACTAGTTCACCGTTTGTTTGCT
HI9920    AACTAGTTTGTGAATATCGTTCCTTA-----GACTAGTTCACCGTTTGTTTGCT
HI9938    AACTAGTTTGTGAATATCGTTCCTTATATAA----AGTTCACCGTTTGTTTGCT
HI9948    AACTAGTTTGTGAATATCGTTCCTTATATAA----AGTTCACCGTTTGTTTGCT
HI9950    AACTAGTTTGTGAATATCGTTCCTTA-----TACTAGTTCACCGTTTGTTTGCT
D. virilis AACTAGTTTGTGAATATCGCTCCTTA-----TACTAGTTCACCGTTTGTTTGCT
```

Figure 4.1: Example of a simple deletion and a complex indel event *D. americana* (*mew* intron 5).

4.3 Results

We summarise simple indel patterns for each intron in table 4.1 and table 4.2 for *D. americana* and *D. simulans* respectively. The dataset for both complex and simple indel events in *D. americana* is shown in Appendix 4.1. First we counted the number of insertions and deletions observed, then we examined the size of insertions and deletions and finally we investigated their frequencies.

Table 4.1: Characteristics of insertions and deletions in all genes for *D. americana*, ordered by their position on the *D. virilis* X chromosome The averages are different from the values in the text because values in the table are averages of mean values for each intron.

		Deletions			Insertions		
Gene	Intron	Number	Mean size (bp)	Mean freq	Number	Mean size (bp)	Mean freq
sol	1	2	1.5	0.077	2	1	0.192
sol	4	1	6	0.071	1	1	0.071
eIF5	2	2	8	0.192	0	NA	NA
shibire	6	1	2	0.143	0	NA	NA
sog	4	5	10.8	0.286	1	1	0.929
mew	5	3	6.33	0.071	1	6	0.071
l(1)1Bi	1	3	7	0.429	2	3	0.643
cv	1	2	1.5	0.071	1	7	0.071
rad	2	2	10.5	0.308	0	NA	NA
GJ16746	1	9	3.78	0.378	1	1	0.071
GJ16746	2	0	NA	NA	1	1	0.071
17292	1	3	1.33	0.119	1	1	0.143
Cyp28c1	1	3	2.33	0.333	2	1.5	0.071
Cyp28c1	2	3	6.67	0.167	1	1	0.786
phl	4	1	3	0.071	1	50	0.071
si	6	5	7.6	0.157	0	NA	NA
hep	1	2	5.5	0.179	0	NA	NA
csw	1	2	8.5	0.321	0	NA	NA
csw	2	1	1	0.143	0	NA	NA
Pros28.1	1	1	1	0.214	0	NA	NA
Yp1	2	0	NA	NA	1	1	0.071
Total / Average (SE)		51	4.97 (0.75)	0.20 (0.026)	17	5.46 (3.73)	0.25 (0.087)

Table 4.2: Characteristics of insertions and deletions in each gene for *D. simulans*. The averages are different from the values in the text because they are averages of mean values for each intron.

Gene	Intron	Deletions			Insertions		
		Number	Mean size (bp)	Mean freq	Number	Mean size (bp)	Mean freq
CG3665	9	2	7.5	0.045	0	NA	NA
CG3595	1	5	3.2	0.226	2	1	0.478
CG32732	1	3	10.33	0.045	1	3	0.318
CG11387	1	1	1	0.2	0	NA	NA
CG1689	4	1	1	0.05	0	NA	NA
CG32688	1	1	26	0.35	0	NA	NA
CG9355	1	8	7.75	0.069	3	1.67	0.067
CG12244	4	1	14	0.727	3	3.33	0.045
CG9533	6	1	1	0.045	1	7	0.364
CG2662	2	5	5.2	0.139	1	5	0.043
CG4420	1	4	4.25	0.239	1	1	0.045
CG11105	3	1	8	0.091	0	NA	NA
CG14045	2	2	8	0.042	0	NA	NA
CG14435	1	1	5	0.227	2	4.5	0.091
CG32683	2	1	9	0.043	0	NA	NA
Crag	9	4	2.5	0.043	1	1	0.043
Cyp4g15	3	3	5.67	0.1	0	NA	NA
Dsor1	1	3	11.67	0.303	2	2.5	0.045
HDAC6	5	2	3.5	0.065	2	2	0.043
Idgf4	1	3	1	0.217	4	6.5	0.076
NetB	2	3	2	0.058	0	NA	NA
para	3	3	9.67	0.043	0	NA	NA
Ptp10D	1	3	15	0.045	1	1	0.045
up	7	3	2	0.347	1	1	0.042
Total / Average (SE)		64	6.84 (1.19)	0.157 (0.033)	25	2.89 (0.56)	0.125 (0.039)

As previously documented in *Drosophila* species, we find a high polymorphism deletion bias in both species (*D. americana*: PDB=3; *D. simulans*: PDB=2.56).

4.3.1 Indel sizes

There is a majority of small insertions and deletions in both *D. americana* and *D. simulans*. 70% of insertions and 63% of deletions in *D. simulans* are 5bp or less. 82% of insertions and 67% of deletions in *D. americana* are 5bp or less. The mean length for deletions is 5.5bp (+/-0.85bp SE) in *D. americana* and 6.3bp (+/-0.97bp SE) in *D. simulans*. The mean length for insertions is 4.8bp (+/-2.86bp SE) in *D. americana* and

3.2bp (+/-0.50bp SE) in *D. simulans*. In both species, insertions appear to be shorter than deletions. This is significant in *D. americana* (Wilcoxon test: p=0.035) but not in *D. simulans* (Wilcoxon test: p=0.280).

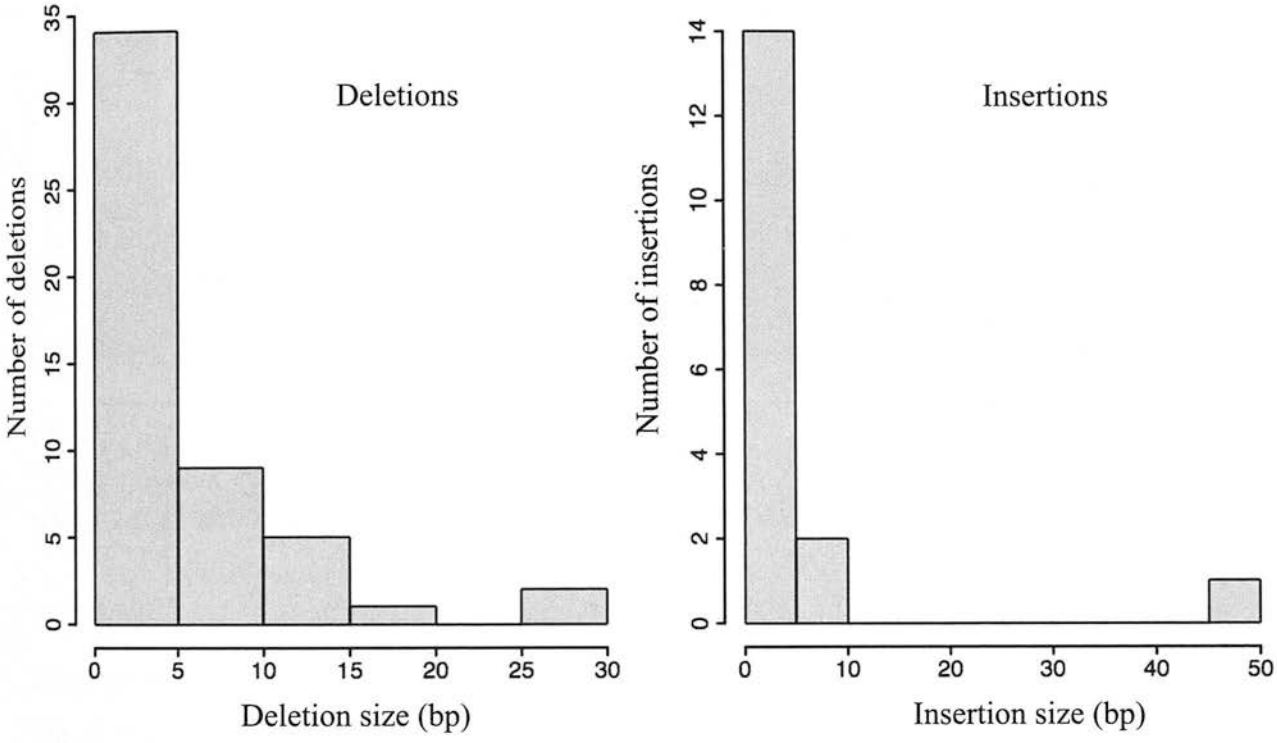

Figure 4.2: *D. americana* deletions and insertions: histogram of sizes

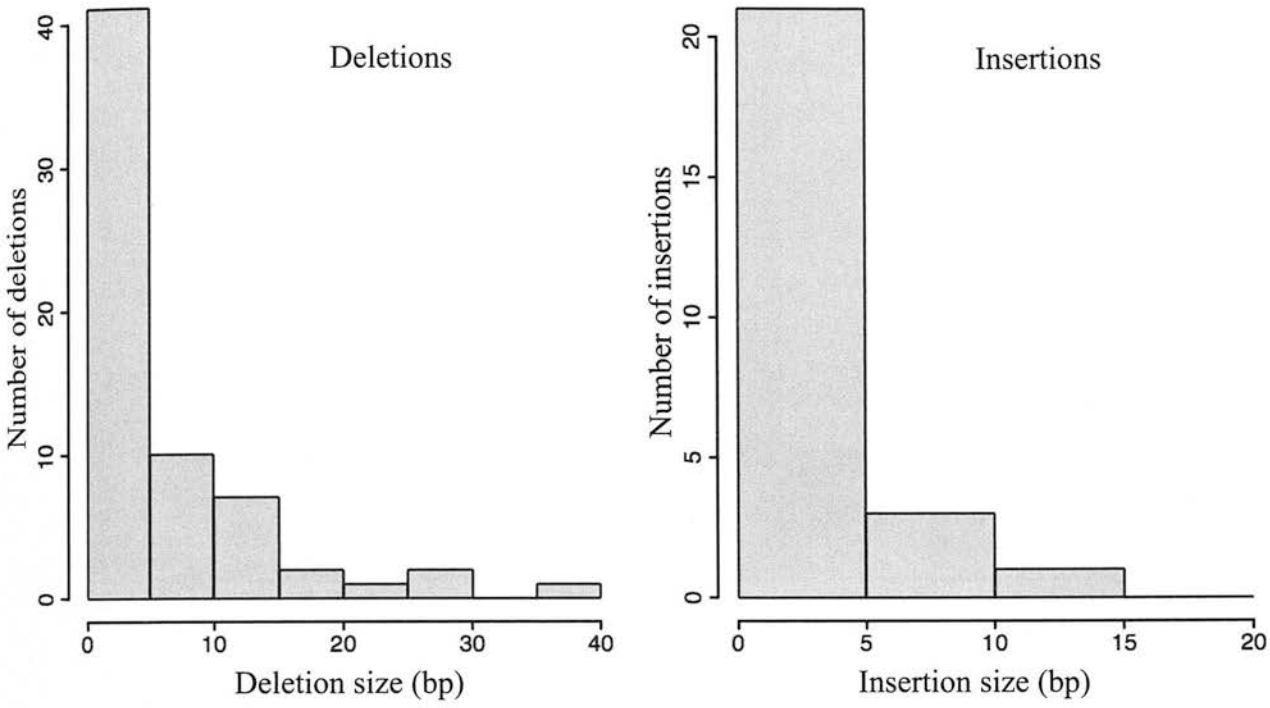

Figure 4.3: *D. simulans* deletions and insertions: histogram of sizes

4.3.2 Indel frequencies

The mean frequency for deletions is 0.23 (+/-0.030 SE) in *D. americana* and 0.14 (+/-0.025) in *D. simulans*. The mean frequency for insertions is 0.26 (+/-0.072 SE) in *D. americana* and 0.11 (+/-0.037) in *D. simulans*. The differences in frequency between insertions and deletions are small and not significant for each species.

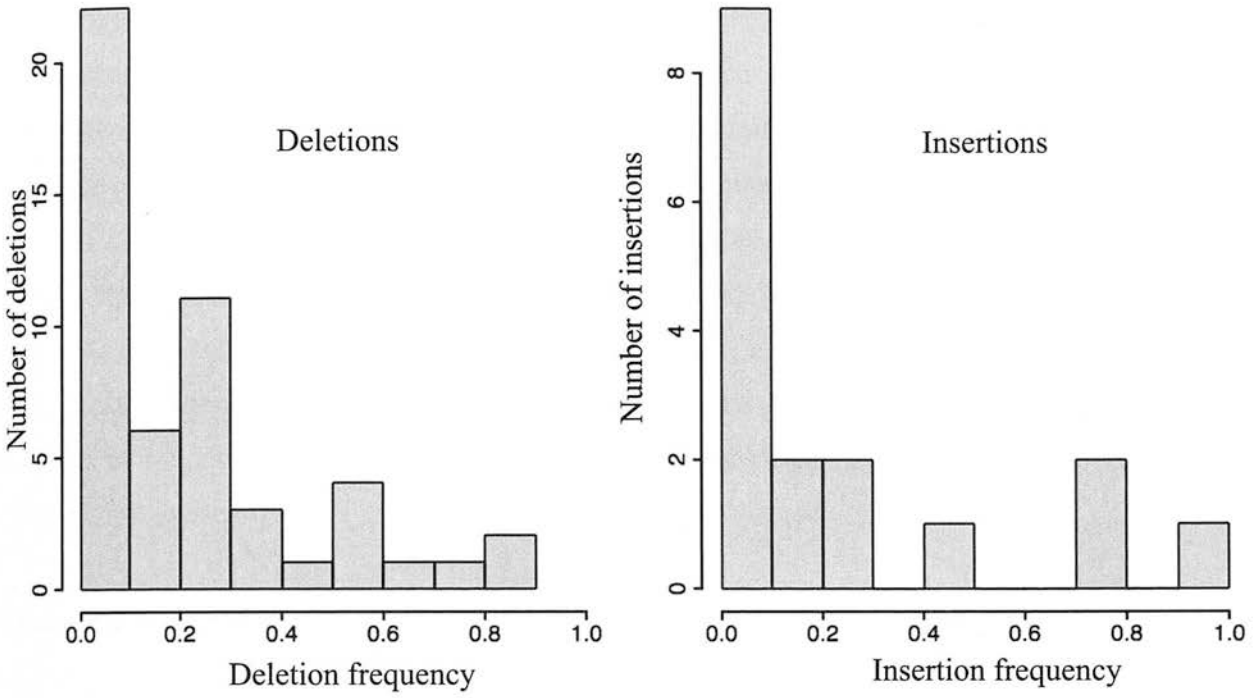

Figure 4.4: *D. americana* deletions and insertions: histogram of frequencies

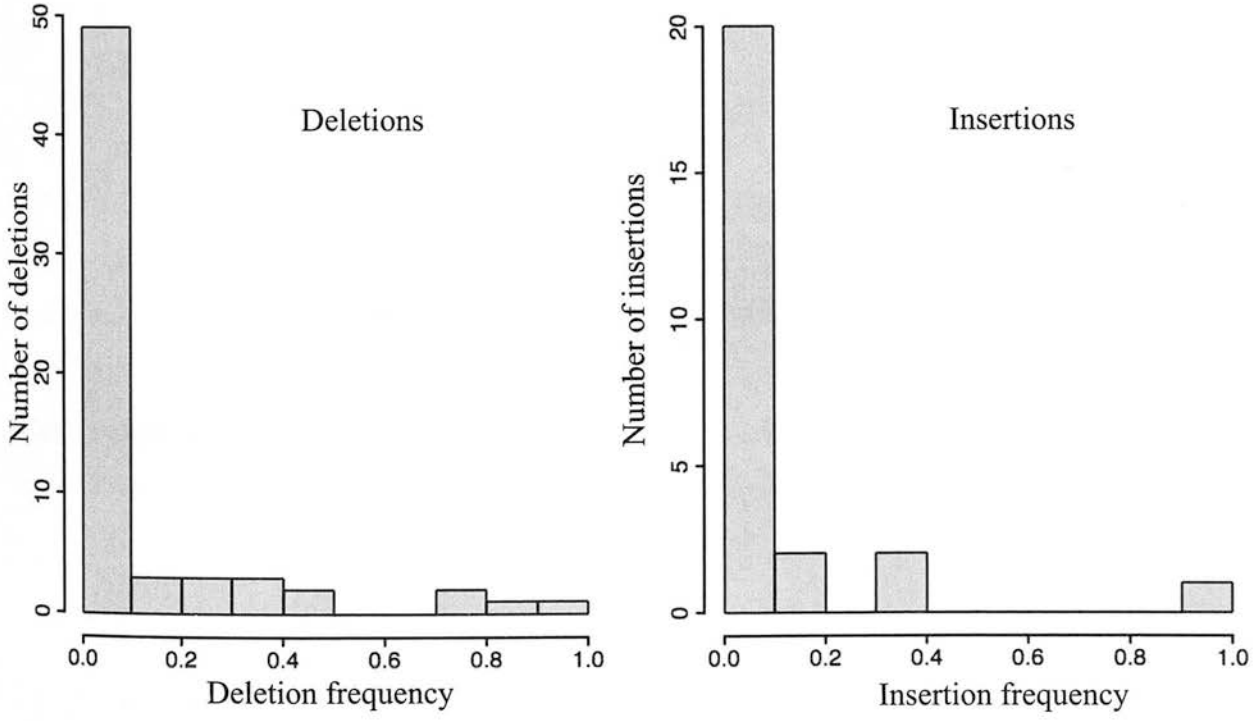

Figure 4.5: *D. simulans* deletions and insertions: histogram of frequencies

4.3.3 Demographic analysis

As mentioned above, changes in population size can affect patterns of sequence evolution and therefore interfere with the ability to detect natural selection. We analysed two datasets (Table 4.3) to determine whether the demographic history of *D. americana* has indeed been stable: one dataset included 20 X-linked coding sequences encompassing 3,665 bp, in which we looked at preferred (P) and unpreferred (U) mutations and the other dataset included 32 X-linked introns encompassing 9,298 bp, in which we looked at GC and AT mutations. The negative values for Tajima's *D* for each dataset indicate an excess of low frequency alleles, consistent with either strong purifying selection or a population expansion.

Table 4.3: Polymorphism indices of the intron and coding sequence dataset used to determine the demographic history of *D. americana*. *G* is the number of loci, *L* is the total number of sites, *n* is the average sample size, *S* is the total number of segregating sites.

	G	*L*	*n*	*S*	π	θ_W	Tajima's *D*
Introns (GC/AT polymorphisms)	32	9,298	13.77	563	0.0131	0.0157	-0.5734
Synonymous sites (P/U polymorphisms)	20	3,665	14.46	120	0.0084	0.0109	-0.8262

In *D. americana*, the method of Zeng and Charlesworth (2009) indicates a recent 3.88-fold increase in population size (Table 4.4). The time since the event is $\tau = 0.25$, where time is a measure in units of the current population size. When we test the difference in ln*L* between the model L_1 (population expansion) and the model L_0 (stable population), we find a significant difference ($\chi^2 = 75.52$, d.f. = 2, $p < 10^{-16}$), suggesting that a model including parameters for a recent population expansion is significantly more likely than a model considering a stable demographic history for *D. americana*.

Table 4.4: Parameters of the demographic models L_0 (stable population) and L_1 (recent population expansion) for *D. americana*. g is the factor of population expansion, τ is the time since the expansion (in units of t/N_2, where t is the number of generations since the population expansion and N_2 is the new population size) and κ is the mutation bias for GC to AT in introns and for U to P codons in coding sequences.

Model	$g(=N_2/N_1)$	$\tau(=t/N_2)$	$\gamma_{cod}(=2N_1s_{cod})$	$\theta_{cod}(=2N_1\mu_{cod})$	κ_{cod}	γ_{int}	θ_{int}	κ_{int}	ln$L$
L_1	3.88	0.25	1.51	0.0033	3.73	0.35	0.0086	2.27	-12349.55
L_0	—	—	1.89	0.0041	5.31	0.42	0.0133	2.41	-12387.31

In *D. simulans*, the method of Zeng and Charlesworth (2009) indicates a recent 15.2-fold increase in population size (Table 4.5). The time since the event is $\tau = 39$, where time is a measure in units of the current population size. When we test the difference in ln$L$ between the model L_1 (population expansion) and the model L_0 (stable population), we find a significant difference (χ^2 =2419.96, d.f. = 2, $p < 0.0001$), suggesting that a model including parameters for a recent population expansion is significantly more likely than a model considering a stable demographic history for *D. simulans*.

Table 4.5: Parameters of the demographic models L_0 (stable population) and L_1 (recent population expansion) for *D. simulans*. g is the factor of population expansion, τ is the time since the expansion (in units of t/N_2, where t is the number of generations since the population expansion and N_2 is the new population size) and κ is the mutation bias for GC to AT in introns and for U to P codons in coding sequences.

Model	$g(=N_2/N_1)$	$\tau(=t/N_2)$	$\gamma_{cod}(=2N_1s_{cod})$	$\theta_{cod}(=2N_1\mu_{cod})$	κ_{cod}	γ_{int}	θ_{int}	κ_{int}	ln$L$
L_1	15.2	0.13	1.42	0.0057	2.53	0.44	0.0031	2.22	-24042.99
L_0	—	—	1.69	0.0144	3.06	0.39	0.0116	2.10	-25252.97

4.3.4 Estimating γ for indels

Using the model L_1 with recent population expansion described above (Table 4.4), we found no evidence for selection for either insertions or deletions in *D. americana*. The selection coefficients for insertions and deletions were not statistically different from 0 ($p > 0.44$), which may be due to insufficient data. In our analysis, a negative value for γ indicates positive selection, whereas a positive value suggests purifying selection. The

estimates for γ are -2.04 for insertions and -0.33 for deletions, suggesting that there may be some positive selection for insertions.

If we assume that the *D. americana* population is in equilibrium using the model L_0 described above, the results are quite different. The estimates for γ are 1.30 (not statistically different from 0, p=0.53) for insertions and 2.08 (nearly significant, p=0.07) for deletions. This would suggest that deletions are more advantageous than insertions. This result illustrates the fact that taking into account the demographic parameters is necessary for unbiased tests of natural selection in non-equilibrium populations.

Similarly for *D. simulans*, using the model L_1 with recent population expansion described above (Table 4.5), we found no evidence for selection for either insertions or deletions. The selection coefficients for insertions and deletions were not statistically different from 0 ($p > 0.55$), which may be due to insufficient data. The estimates for γ for *D. simulans* are 0.83 for insertions and -0.48 for deletions.

If we assume that the *D. simulans* population is in equilibrium using the model L_0 described above, the results are quite different. The estimates for γ are 6.40 (statistically different from 0, $p=2.99\times10^{-4}$) for insertions and 4.64 (significant, $p=6.25\times10^{-6}$) for deletions. This would suggest that both deletions and insertions are subject to purifying selection. This result confirms the importance of taking into account the demographic parameters when testing for natural selection.

4.4 Discussion

The ratio of number of deletion polymorphisms to the number of insertions polymorphisms is a close estimate of the mutational pattern unless relatively strong selection is acting. As previously found in *Drosophila* (Petrov et al. 1996, Petrov and Hartl 1998; Comeron and Kreitman 2000; Ometto et al. 2005; Presgraves 2006), we observed an excess of deletions in both *D. americana* and *D. simulans*. Our values of 3 and 2.56 for polymorphism deletion bias are high when compared with the estimate of 1.35 from a genome-wide study of *D. melanogaster* introns (Comeron and Kreitman,

2000), but considerably lower than estimates of deletion bias for paralogous divergence from studies on *Helena* retrotransposons in *D. melanogaster* and *D. virilis* (Petrov et al. 1996, Petrov and Hartl 1998). This has been attributed to constraint associated with splicing mechanisms in introns (Ptak and Petrov 2002) or a possible adaptive role for indels in *Helena* retrotransposons (Charlesworth 1996).

As suggested by the negative correlation between intron length and recombination rate (Comeron and Kreitman 2000), recombination can affect natural selection on insertions and deletions. This paper found that the polymorphic deletion bias did not vary significantly with recombination rate, suggesting that the correlation is not due to a change in mutational deletion to insertion bias with recombination.

4.4.1 Indel sizes

We found that most insertions and deletion sizes were shorter than 5bp, consistent with other *Drosophila* studies (Comeron and Kreitman 2000; Parsch 2003; Ometto et al. 2005). Gregory (2004) suggested that the shorter indels may be due to replication slippage, which is more frequent than other indel processes such as transposable elements or microsatellites. Deletions are also longer than insertions in both species. This suggests that introns should evolve to be shorter, which is not the case: intron lengths are relatively stable over evolutionary time (Stephan et al. 1994). The fact that most deletions are very short could mean that they are only slightly deleterious and could be fixed by genetic drift. Natural selection to restore intron length may then occur, fixing a large insertion to compensate for the DNA loss through small deletions (Parsch 2003).

4.4.2 Indel frequencies

Several intron evolution models argue that the DNA loss through the high deletion rate is compensated by higher segregating frequencies of insertions due to positive selection, leading to higher probabilities of fixation (Comeron and Kreitman 2000; Parsch 2003). We did not find a significant difference in frequencies of insertions vs. deletions for either *D. americana* or *D. simulans*, even though it is worth noting that the mean frequency for insertions is higher than that for deletions in *D. americana*. Comeron and

Kreitman (2000) found elevated frequencies of insertions relative to deletions in regions of high recombination, suggesting that this higher mean frequency of insertions is indeed due to natural selection. In a study using three *Drosophila* species, which allows the polarisation of fixed indels, Presgraves (2006) observed higher frequencies for small insertions than for deletions, particularly on the X chromosome. The polymorphism deletion bias he observed was not different between the X and autosomes, indicating that the higher probabilities of fixation of insertions on the X are not due to a mutational bias in *D. melanogaster*. Since the genes studied in this chapter are all located on the X chromosome, we cannot compare our results with values for autosomes. The main explanation for higher probabilities of fixation of insertions on the X is the higher rate of crossing over on the X. This would increase the efficacy of natural selection relative to autosomes if insertions were indeed favoured (Betancourt and Presgraves, 2002), and it would also enhance transmission of insertions through a biased gene conversion-gap repair process (Marais 2003).

4.4.3 Demographic analysis

A recent change in demographic history can interfere with the ability to detect the effects of natural selection on DNA sequences. The results of the demographic analysis for *D. americana* further shows that it is essential to test for the demographic history of species before attempting to detect natural selection. Overlooking a population size change may well bias the results and give wrong estimates of natural selection. Schaeffer (2002) also considered demographic history prior to analysing insertions and deletions in *D. pseudoobscura*. He estimated the value of Nr, the product of the current effective population size and the exponential growth rate (Slatkin and Hudson 1991). His estimate of $Nr = 7$ shows that the negative Tajima's D found for *Adh* coding sites was actually due to population expansion and not purifying selection (Schaeffer et al. 2001; Schaeffer 2002).

4.4.4 Estimating γ for indels

Despite the trend of higher insertion frequencies than deletion frequencies, our results with the maximum likelihood method used here, which is a modified version of the method by Zeng and Charlesworth (2009), do not suggest significant positive selection for insertions. The γ estimate is positive, however, suggesting that the lack of significance may only be due to insufficient data and that insertions may be selectively favoured as in models of indel evolution proposed by Comeron and Kreitman (2000) and Parsch (2003).

The γ estimate for deletions is very close to 0, suggesting that deletions are close to neutral, which is compatible with most models of indel evolution. This might also indicate that the deletions we observed, which were mostly small deletions, did not interfere with splicing mechanisms in introns (Ptak and Petrov 2002).

Why did this study fail to find a difference between insertions and deletions, while other studies have found this difference? One possibility is that the difference is real, but that the present study lacks power. Presgraves (2006) used 148 published sequences for introns from 68 genes in *D. melanogaster*, so his dataset was more substantial than ours, although it is a mixture of data from various sources, and he observed 135 polymorphic indel events. Ometto et al. (2005) used 22 intergenic regions and 54 introns in *D. melanogaster* and observed 93 non-repetitive polymorphic indel events. In the *D. americana* dataset, we only had 21 introns from 17 genes and we observed 68 simple polymorphic indel events. In the *D. simulans* dataset, we had 24 introns from 24 genes and we observed 89 simple polymorphic indel events. Therefore, the data sets of Presgraves (2006) and Ometto et al. (2005) were approximately 7 and 3.6 times bigger than the present data set in terms of number of fragments studied.

Another possibility is that the differences found are due to difference in alignment methods. In this study, an initial analysis (not shown) was conducted on a data set aligned by hand. Using the hand alignments, there was a significant elevation in the frequency of insertions, which disappeared when the alignments were instead conducted with MCALIGN2 (Wang et al. 2006). The sequences from Presgraves (2006) were aligned using a different method, DIALIGN2 (Morgenstern 1999) followed by hand alignment, which might cause the observed difference in insertion frequencies. Furthermore, the original version of MCALIGN (Keightley and Johnson. 2004) was shown to perform

better than DIALIGN. However, this requires that the difference in alignment methods only be a problem on the X chromosome.

Finally, the differences in insertion frequencies might be due to the different species used. This seems likely as only one species (*D. melanogaster*) clearly showed favoured insertions in Presgraves (2006), and this only on the X.

5 Selection for codon usage bias and on GC polymorphisms and recombination.

Contributing authors:

- I collected the data and performed the analyses.
- K. Zeng performed the Maximum-Likelihood analyses.
- A. Betancourt helped with recombination analyses.
- B. Charlesworth advised on the project.

5.1 Introduction

Synonymous sites, as well as non-coding sequences, have often been used as neutrally evolving standards in tests for natural selection. However, there is evidence that these sites can also be subject to natural selection, making it harder to find neutrally evolving sites. In this chapter, we investigate patterns of natural selection on these putatively neutrally evolving sites, and the influence of recombination on these patterns.

There is a large amount of evidence showing that synonymous sites are subject to selection for codon usage bias (Ikemura 1981; Ikemura 1985; Sharp and Li 1987; Akashi 1995). Some codons are indeed favoured over other synonymous codons and species have different sets of preferred codons. Preferred codons often correspond to the most abundant tRNA for each amino acid (Ikemura 1981; Shields et al. 1988; Powell and Moriyama 1997). Codon usage bias is positively correlated with gene expression (Post and Nomura 1980; Ikemura 1981; Gouy and Gautier 1982; Sharp and Li 1986; Shields et al. 1988; Powell and Moriyama 1997; Duret and Mouchiroud 1999). Hypotheses to explain these correlations include selection for translation efficiency in terms of translation accuracy and speed (Kurland 1987; Bulmer 1991; Akashi 1994; Moriyama and Powell 1998). Several population genetics studies have aimed at estimating the

intensity of selection for translational efficiency (Hartl et al. 1994; Cutter and Charlesworth 2006; Plotkin et al. 2006; Zeng and Charlesworth 2009; dos Reis and Wernisch 2009).

It has also been shown that another non-neutral process, biased gene conversion, can affect non-coding sequences in similar ways to natural selection (Gutz and Leslie 1976; Nagylaki 1983; Marais 2003). Biased gene conversion occurs during the repair of double strand breaks during recombination (Galtier et al. 2001; Marais 2003; Galtier et al. 2006; Galtier and Duret 2007; Galtier et al. 2009), involving the repair of mismatches by preferentially replacing the erroneous base by a G or a C to match the opposite strand rather than the other way around.

Recombination is strongly associated with both selective and mutational processes. Low recombination can reduce the efficacy of natural selection (Hill and Robertson 1966), so that a positive correlation between recombination rate and codon usage bias is expected and has been found in *Drosophila* (Comeron et al. 1999; Haddrill et al. 2007). It has been found that both gene length and recombination rate affect codon usage bias (Moriyama and Powell 1998; Comeron et al. 1999; Duret and Mouchiroud 1999). Gene conversion should also be affected by the recombination rate because it only occurs to correct mismatches in heteroduplexes. This may explain the observed reduced GC content in low-recombination regions of the genome (Singh et al. 2005a; Díaz-Castillo and Golic 2007; Haddrill et al. 2007). Since biased gene conversion is a non-selective process, it should affect coding and linked non-coding sequences in a similar way (Marais et al. 2003). However, this assumes that recombination rates are the same in exons and linked introns. Differences in recombination rates between coding sequences and introns might affect the results of studies aiming at distinguishing between biased gene conversion and natural selection (Marais et al. 2003). Therefore, we have also investigated the difference in recombination rate estimates between exons and introns.

It is interesting to note that both selection for codon usage bias and biased gene conversion affect the base composition of DNA sequences (Li 1987; Bulmer 1991). Preferred codons in *Drosophila* generally end in G or C; therefore both processes tend to increase the GC content over evolutionary time, leading to a positive correlation between

GC content and recombination rate in non-coding and synonymous sites (Marais et al. 2001). This positive correlation has been used by Marais et al. (2003) to determine whether biased gene conversion or Hill-Robertson interference was causing the base composition. Indeed, if the correlation between GC content and recombination is only found in synonymous sites, then it is likely to be due to selection for codon usage only, whereas if the correlation is also observed in noncoding sites, then a non-selective force, such as biased gene conversion, is probably responsible. In opposition to these GC-biased processes, a mutational bias towards A or T nucleotides has been well documented in *Drosophila* (Kliman and Hey 1994; Singh et al. 2007; and also in direct mutational studies: Haag-Liautard et al. 2007; Keightley et al. 2009).

5.2 Materials and methods

5.2.1 Codon usage bias

We polarised synonymous polymorphisms in the *D. americana* coding sequences dataset (Chapter 3) with the *D. virilis* sequence. The final dataset included 26 coding sequences for which there were synonymous changes, including 10 autosomal loci from Maside and Charlesworth (2007). Using a codon preference table for *D. virilis* (Betancourt et al. 2009) (Appendix 5.1), we assigned preferred codons (P) and unpreferred codons (U) in each species and then determined if the synonymous site change within *D. americana* was P > P, U > U, P> U or U > P. It is expected that P > P and U > U changes are neutral, P > U changes are deleterious, and U > P changes are advantageous (Bulmer 1991; Akashi 1995). We obtained the frequencies and numbers of synonymous site changes in each category.

As described in Chapter 4, we used a maximum likelihood method (Zeng and Charlesworth 2009; Zeng and Charlesworth 2010) to estimate the strength of selection on U > P polymorphisms, after taking into account a recent population expansion in *D. americana*. The models use both coding and non-coding sequences. We compared a model with population expansion and selection for AT > GC polymorphisms in introns

(L_2) with the same model that also included selection for U > P polymorphisms in coding sequences (L_1) and tested the difference between the two models using a likelihood ratio test.

5.2.2 AT/GC polymorphisms

Similarly, we obtained the counts and frequencies of AT > AT, GC > GC, GC > AT and AT > GC polymorphic changes for each intron in the *D. americana* intron dataset, as described in Chapter 3.

Again, we used a maximum likelihood method (Zeng and Charlesworth 2009; Zeng and Charlesworth 2010) to estimate the strength of selection on AT > GC polymorphisms in introns, after taking into account a recent population expansion in *D. americana*. We compared a model with population expansion and selection for U > P polymorphisms in coding sequences (L_3) with the same model that also included selection for AT > GC polymorphisms in introns (L_1) and tested the difference between the two models using a likelihood ratio test.

5.2.3 Recombination

A composite likelihood estimation method (Hudson 2001) is implemented in the program LDhat 2.1 (McVean et al. 2002) to estimate the amount of population recombination r in the history of a set of aligned sequences. The method assumes a constant mutation rate and, for this reason, analysis is restricted to polymorphic sites with two different segregating sites. We estimated recombination rates $4N_e r$, where N_e is the effective population size and r is is the genetic map distance across the region analysed (the product of the physical distance and the per site rate of recombination across the region) for each exon and each intron in the *D. americana* polymorphism dataset via the *pairwise* program in LDhat 2.1. The dataset comprised 28 introns and 20 exons from 17 genes. We then divided the estimate by the number of sites in each alignment to obtain an estimate per bp. We further scaled it by dividing by θ_W, ($4N_e\mu$) as in Andolfatto and Przeworski (2000), which should make this ratio (r/μ) independent of effective population size under the standard neutral model assumptions. Scaling to θ_W could also be important if

background selection is reducing the effective population size for some loci (Charlesworth et al. 1995). The ratio is then an estimate of the number of recombination events per mutation.

5.3 Results

5.3.1 Codon usage bias

There are nearly three times as many P>U changes as U>P changes (162 vs. 56; Table 5.1). We examined the frequency distributions of polymorphisms in each of these categories and each of the non-coding DNA classes (Figure 5.1). Figure 5.1 shows that selection to maintain optimal codon usage is likely to account for a considerable fraction of the overall skew toward low-frequency variants in the frequency spectrum of synonymous polymorphisms. Following Haddrill et al. (2008a), we consider P > P and U > U classes (accounting for ~28% of our polymorphisms) as a neutral frame of reference. P > U changes (accounting for ~53% of our polymorphisms) are putatively negatively selected and, consistent with this expectation, this category shows the strongest skew toward rare variants and seems significantly more negatively skewed than the P > P / U > U class. In contrast, a greater proportion of U > P changes, the putatively positively selected class and ~19% of our polymorphisms, are seen at intermediate or high frequency. The mean frequency for U > P changes was significantly higher than that for both P > U changes (Wilcoxon test, p=0.022) and pooled P > P and U > U changes (Wilcoxon test, p=0.030).

Table 5.1: Counts and frequencies of P > U and U > P polymorphisms. Averages are calculated gene by gene. We only observed 31 U > U polymorphisms and 13 P > P polymorphisms. n is the number of lines used for each gene.

Gene	n	P>U	MeanFreqP>U	U>P	MeanFreqU>P
anon-66Db	5	8	0.275	0	NA
Cdc37	5	5	0.2	1	0.2
cp36	5	1	0.2	0	NA
csw	14	7	0.122	4	0.143
cv	14	2	0.071	3	0.714
Cyp28c1	14	10	0.157	1	0.143
Ddx1	6	9	0.204	2	0.75
dor	14	2	0.393	0	NA
dos	6	15	0.244	6	0.5
eIF5	12	15	0.194	4	0.542
elav	50	13	0.035	2	0.29
ful	5	1	0.4	0	NA
GJ16746	14	3	0.405	1	0.071
hep	14	1	0.143	1	0.643
Kni	5	3	0.267	3	0.733
l(1)1Bi	14	6	0.321	1	0.929
msl3	5	1	0.2	4	0.65
Per	5	6	0.233	2	0.4
Pros28.1	14	3	0.167	2	0.107
rh4	5	6	0.367	2	0.3
Sina	5	4	0.3	0	NA
sol	13	2	0.462	2	0.385
su(Hw)	5	6	0.233	5	0.36
su(s)	5	6	0.3	1	0.2
Tll	8	19	0.303	3	0.292
Yp1	13	8	0.212	6	0.244
Total / Average (SE)		162	0.246 (0.021)	56	0.409 (0.054)

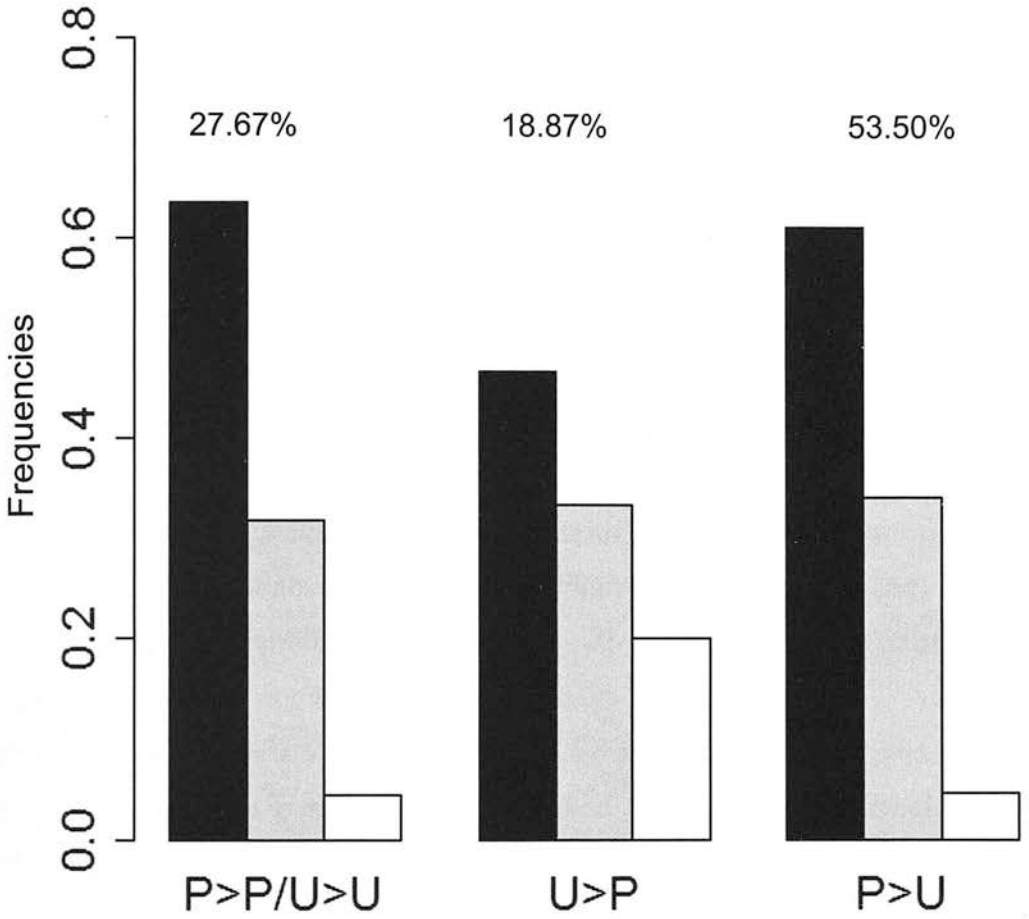

Figure 5.1: The distribution of frequency classes of polymorphisms for different types of synonymous site changes. P > P / U > U includes preferred to preferred and unpreferred to unpreferred changes, P > U are preferred to unpreferred changes, and U > P are unpreferred to preferred changes. The low-frequency class (black) includes polymorphisms at a frequency lower than 0.2, the intermediate frequency class (grey) includes polymorphisms at a frequency between 0.2 and 0.8, and the high frequency class (white) include polymorphisms at a frequency higher than 0.8. The numbers above each type of synonymous site change indicate the percentage of the total synonymous polymorphisms of each type.

For the likelihood method, we compared two models and tested whether they were significantly different using a likelihood ratio test. When comparing L_1 with L_2 (Table 5.3), the difference was highly significant ($\chi^2 = 29.64$, 1 df, $p<0.0001$). Therefore,

L_1 is more likely than L_2 and selection favouring U > P polymorphisms is necessary to explain the observed data, with a γ estimate of 1.51.

5.3.2 AT to GC polymorphisms

The total number of GC to AT polymorphisms is not significantly different from the total number of AT to GC polymorphisms (χ^2=0.07, p=0.786) (Table 5.2). Since introns with different GC contents may be evolving under different selective constraints for base composition, we separated introns into classes depending on their GC content as in Haddrill and Charlesworth (2008). High, medium and low GC content sequences correspond to loci with GC contents in the ranges 43-56% (n=2, mean 50%), 37-42% (n=15, mean 39%) and 26-36% (n=15, mean 32%). When separating introns according to their GC content, no category has a significant difference between number of GC to AT polymorphisms and number of AT to GC polymorphisms (χ^2 tests, p>0.14).

For each site polymorphic for GC and AT, we counted the number of GC and AT polymorphisms. The number of GC variants in introns with low GC content is not significantly different from the number of AT variants (χ^2=0.80, p=0.373) but it is significantly higher in introns with medium and high GC content (medium: χ^2=53.33, p<0.0001; χ^2=19.59, p<0.0001).

The mean frequency of GC to AT polymorphisms is not significantly lower than the mean frequency of AT to GC polymorphisms (t = -1.99, df = 52.87, p-value = 0.052). When separating introns according to their GC content, no category has a significant difference between mean frequency of GC to AT polymorphisms and the mean frequency of AT to GC polymorphisms (p>0.062). The difference is closest to significance in the medium GC content category of introns, with the mean frequency of GC to AT polymorphisms being lower than the mean frequency of AT to GC polymorphisms.

The mean frequency for AT to GC polymorphisms in introns with high GC content is not significantly different from the mean frequency of AT to GC polymorphisms in introns with low GC content (t = -0.15, df = 1.16, p-value = 0.901) or introns with medium GC content (t = 0.20, df = 1.03, p-value = 0.873). The mean frequency for GC to AT polymorphisms in introns with high GC content is not

significantly different from the mean frequency in GC to AT polymorphisms in introns with low GC content (t = -1.58, df = 11.84, p-value = 0.141) or introns with medium GC content (t = -2.26, df = 2.78, p-value = 0.116).

Table 5.2: GC > AT and AT > GC polymorphisms counts and mean frequencies for all introns, ordered by their position on the *D. virilis* X chromosome. The total number of GC and AT variants among polymorphic sites for AT/GC is also given.

Gene	Intron	GC	GC>AT	AT>GC	Prop GC>AT	MeanFreq GC>AT	MeanFreq AT>GC	No. GC	No. AT
dor	1	0.3	0	0	NA	NA	NA	0	0
dor	2	0.328	0	6	0	NA	0.25	21	63
sol	1	0.382	3	5	0.375	0.103	0.231	50	54
sol	2	0.385	1	0	1	0.083	NA	11	1
sol	4	0.377	24	22	0.522	0.234	0.232	306	296
eIF5	1	0.342	3	4	0.429	0.444	0.375	38	46
eIF5	2	0.349	12	11	0.522	0.240	0.281	145	134
shibire	6	0.399	24	13	0.649	0.215	0.455	337	164
sog	4	0.366	15	24	0.385	0.305	0.345	254	273
mew	5	0.345	18	24	0.429	0.224	0.217	253	296
l(1)1Bi	1	0.392	32	18	0.64	0.195	0.259	415	272
l(1)1Bi	2	0.335	0	2	0	NA	0.143	4	24
cv	1	0.383	3	6	0.333	0.071	0.107	48	78
cv	2	0.194	1	1	0.5	0.929	0.429	7	21
rad	1	0.525	5	3	0.625	0.129	0.5	82	30
rad	2	0.353	4	8	0.333	0.096	0.216	66	76
GJ16746	1	0.406	9	9	0.5	0.256	0.197	108	121
GJ16746	2	0.387	1	2	0.333	0.071	0.107	16	26
17292	1	0.398	10	5	0.667	0.271	0.206	116	90
17292	2	0.479	1	1	0.5	0.077	0.077	13	13
Cyp28c1	1	0.346	13	18	0.419	0.082	0.274	236	196
Cyp28c1	2	0.354	15	18	0.455	0.190	0.179	210	244
phl	4	0.423	20	11	0.645	0.121	0.2	254	139
si	6	0.409	11	16	0.407	0.169	0.317	198	179
hep	1	0.404	12	6	0.667	0.263	0.357	153	98
hep	2	0.218	1	2	0.333	0.071	0.321	22	20
csw	1	0.383	0	1	0	NA	0.143	2	12
csw	2	0.429	4	5	0.444	0.146	0.286	66	58
Pros28.1	1	0.316	1	0	1	0.071	NA	13	1
Pros28.1	2	0.317	2	0	1	0.107	NA	25	3
Yp1	1	0.352	1	0	1	0.071	NA	13	1
Yp1	2	0.303	2	1	0.667	0.143	0.857	36	6

The correlation between the GC content at the third codon position in coding sequences (GC3) and the mean intron GC content is not significant (Spearman's rank correlation rho=0.31, p=0.255). This is very similar to the highly significant correlation coefficient of 0.33 found for the *D. virilis* genome (Heger and Ponting 2007), suggesting that the lack of significance here is due to insufficient data.

The correlation between the mean frequency of GC to AT polymorphisms and the intron GC content is not significant (Spearman's rank correlation rho=0.03, p=0.861) (Figure 5.2).

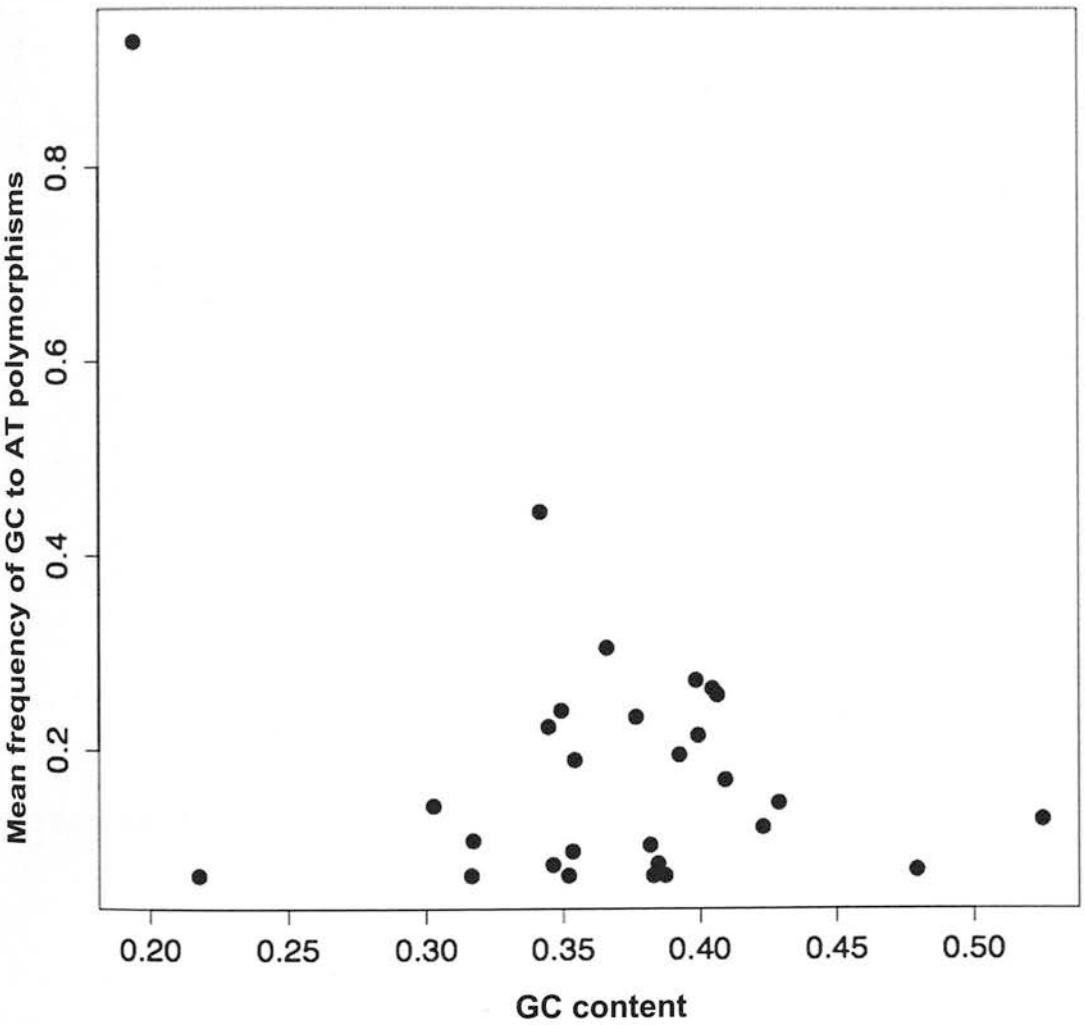

Figure 5.2: Mean frequency of GC to AT polymorphisms against GC content in introns.

The correlation between the proportion of GC to AT polymorphisms and the intron GC content is not significant (Spearman's rank correlation, rho=0.12, p=0.529) (Figure 5.3).

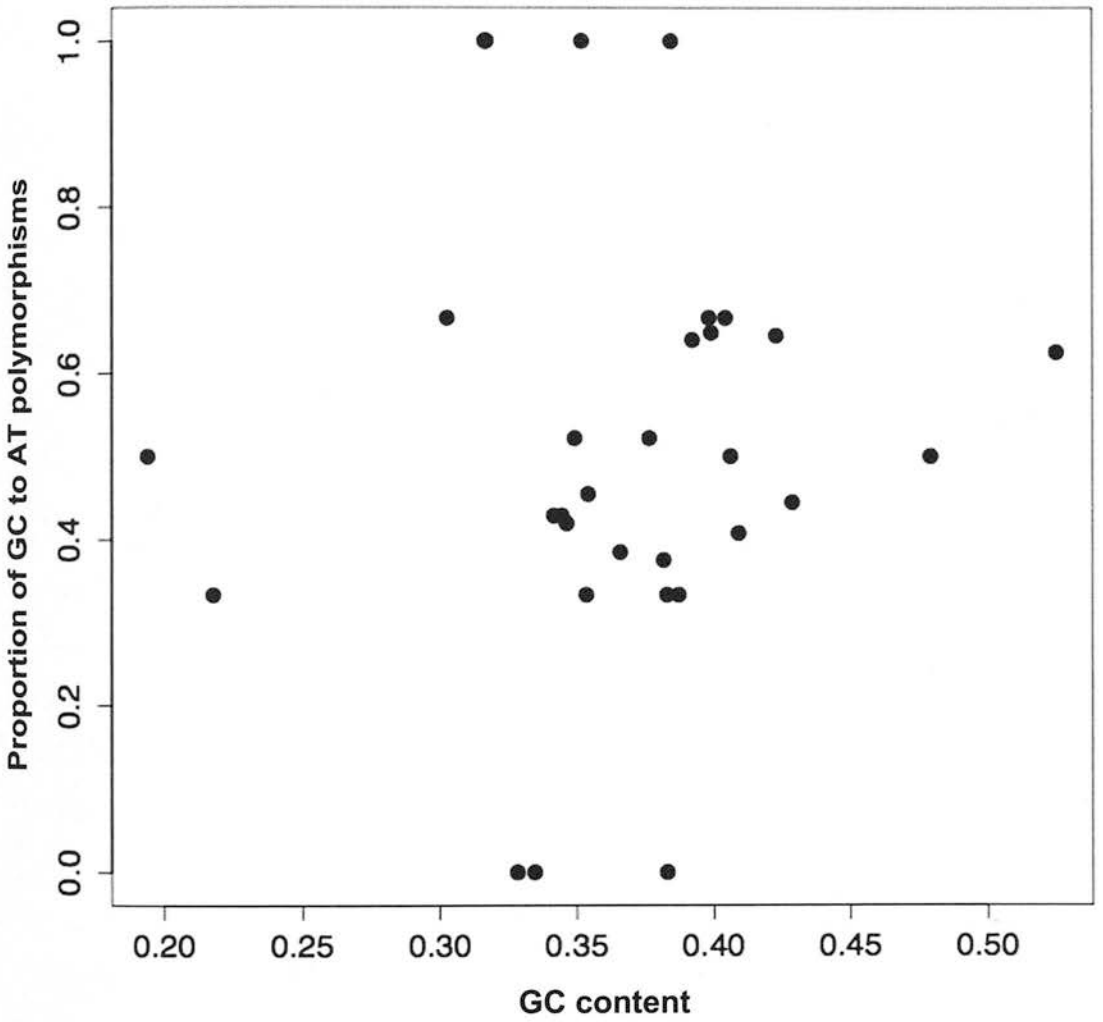

Figure 5.3: Proportion of GC to AT polymorphisms against GC content in introns.

For the likelihood method, we used two models (Table 5.3) and compared them with a likelihood ratio test. When comparing L_1 with L_3, the difference was significant (χ^2 = 7.92, 1 df, p=0.005). Therefore, L_1 is more likely than L_3 and selection for AT > GC polymorphisms is necessary to explain the observed data, with a γ estimate of 0.35.

Table 5.3: Parameters for the different models tested. L_0 is a model with selection on AT > GC and U > P polymorphisms, L_1 is the same model with a population expansion and both selection for U > P polymorphisms in coding sequences and selection for AT > GC polymorphisms in introns, L_2 is a model with a population expansion and only selection for AT > GC polymorphisms in introns and L_3 is a model with a population expansion and only selection for U > P polymorphisms in coding sequences. g is the factor of population expansion, τ is the time since the expansion (in units of), κ is the mutation bias for GC to AT in introns and for U to P codons in coding sequences, and AIC is Akaike's information criterion.

Model	g (=N_2/N_1)	τ (=t/N_2)	γ_{cod} (=$2N_1s_{cod}$)	θ_{cod} (=$4N_1u_{cod}$)	κ_{cod}	γ_{int}	θ_{int}	κ_{int}	ln$L$	AIC
L_0	—	—	1.89	0.0041	5.31	0.42	0.0133	2.41	-12387.31	24786.6
L_1	3.88	0.25	1.51	0.0033	3.73	0.35	0.0086	2.27	-12349.55	24715.1
L_2	4.10	0.24	—	0.0077	0.85	0.35	0.0085	2.26	-12364.37	24742.7
L_3	3.88	0.26	1.51	0.0033	3.74	—	0.0101	1.62	-12353.51	24721.0

5.3.3 Recombination

After excluding values of $4N_er$ of 100, there was no significant difference between the mean recombination rate estimate (r/μ) for introns and that for exons (Wilcoxon test, W = 142, p-value = 0.602) (Table 5.4). The mean recombination rate estimate was 13.05 (4.18 SE) recombination events per mutation for introns and 8.22 (2.13 SE) recombination events per mutation for exons. We also tested the difference in mean recombination rate estimate between introns and exons of the same gene using a paired t-test but it was still not significant (p-value = 0.436), probably due to our small simple size. These estimates appear to be very variable, even within genes.

Table 5.4: Recombination rate estimates for each exon and intron, ordered by their position on the *D. virilis* X chromosome. The maximum value for $4N_er$ estimated by LDhat is 100, so values above 100 are treated as 100. I means introns and E means exon.

Gene	Part	$4N_er$	Alignment length (bp)	$4N_er$ per bp	$4N_er$ per bp scaled by θ_W (r/μ)
dor	I1	4	64	0.063	10.02
dor	E2	4	641	0.006	2.36
dor	I2	8	924	0.009	1.95
dor	E3	4	31	0.129	50.79
eIF5	E2	97	483	0.201	23.82
eIF5	I2	100	640	–	–
shibire	I6	100	833	–	–
sog	I4	100	680	–	–
mew	I5	100	885	–	–
l(1)1Bi	E1	34	88	0.386	39.07
l(1)1Bi	I1	100	913	–	–
l(1)1Bi	E2	50	372	0.134	13.56
l(1)1Bi	I2	32	66	0.485	40.12
cv	I1	44	366	0.120	8.80
cv	E2	20	201	0.100	19.27
cv	I2	36	87	0.414	11.32
cv	E3	20	40	0.500	96.34
rad	I1	80	236	0.339	13.83
rad	I2	53	553	0.096	3.61
GJ16746	I1	100	443	–	–
GJ16746	E2	39	419	0.093	13.78
GJ16746	I2	0	64	0	0
17292	I1	100	879	–	–
17292	I2	100	62	–	–
Cyp28c1	I1	100	608	–	–
Cyp28c1	E2	98	346	0.283	20.72
Cyp28c1	I2	65	473	0.137	3.80
Cyp28c1	E3	6	82	0.073	5.23
phl	I4	100	1267	–	–
si	I6	91	44	2.068	64.16
hep	I1	100	636	–	–
hep	E2	100	391	–	–
hep	I2	10	72	0.143	4.55
csw	E1	43	103	0.417	57.76
csw	I1	53	81	0.654	34.66
csw	E2	79	118	0.669	92.66
csw	I2	10	87	0.115	2.63
csw	E3	100	718	–	–
Pros28.1	E1	100	64	–	–
Pros28.1	I1	4	78	0.051	10.22
Pros28.1	E2	100	230	–	–
Pros28.1	I2	5	68	0.074	6.23
Pros28.1	E3	4	380	0.011	2.95
Yp1	E1	100	123	–	–
Yp1	I1	6	68	0.088	5.04
Yp1	E2	25	383	0.065	8.05
Yp1	I2	2	80	0.025	0.86
Yp1	E3	3	213	0.014	1.73

5.4 Discussion

5.4.1 Codon usage bias

We observed a majority of P > U polymorphisms, suggesting that there is a mutational bias toward unpreferred codons, ending in AT rather than GC, as found in Maside and Charlesworth (2004). There is indeed evidence for an AT-biased mutation process in *Drosophila* (Kliman and Hey 1994; McVean and Vieira 2001; Haag-Liautard et al. 2007; Keightley et al. 2009). Despite this bias, we observed significantly higher frequencies for U > P polymorphisms than for P > U or P > P and U > U. This suggests that preferred codons are being actively favoured by positive selection. The maximum-likelihood method confirms that there is selection for codon usage favouring U to P polymorphisms, after taking into account a recent population expansion (see Chapter 4). The γ estimate of 1.51 is slightly lower than that (2.6) found for a different set of 18 genes, both autosomal and X-linked (Maside and Charlesworth 2004), suggesting that, either autosomal genes have higher codon usage bias than X-linked genes, or that population expansion caused an overestimate of the strength of selection for codon usage bias in the previous study. This pattern could also be observed if biased gene conversion favouring GC is acting on these coding sequences.

5.4.2 AT to GC polymorphisms

Unlike the coding sequence data, we find no significant difference between the number of GC > AT and AT > GC polymorphisms in introns, suggesting that different evolutionary processes affect coding and non-coding sequences. Haddrill and Charlesworth (2008) separated non-coding sequences according to their GC content. If regional variation in GC content is the result of selection or biased gene conversion favouring GC in some regions of the genome more strongly than in other regions, we might expect sequences with high GC content to have the strongest bias towards AT to GC polymorphisms. The correlation between the proportion of GC to AT polymorphisms and the intron GC content is, in fact, positive, as in Haddrill and Charlesworth (2008), though it is not significant. Similarly, if there is stronger selection or biased gene conversion in GC-rich

sequences, there should also be higher frequencies of GC derived variants in GC-rich sequences. Casillas et al. (2007) suggest that the excess of low-frequency GC > AT changes might be due to purifying selection preserving functional GC nucleotides rather than a change in mutation rate or biased gene conversion in CNS regions. However, we do not find any difference between the different GC content categories of introns. The only clear signature of selection or biased gene conversion that we do observe is the significantly higher number of GC variants than the number of AT variants in introns with high GC content.

However, there is some further evidence that GC mutations are favoured. Even though the difference is not statistically significant, AT > GC polymorphisms do have higher frequencies than GC > AT changes. We also see a weak signature of regional variation in GC content (possibly due to biased gene conversion, Marais 2003), as GC content at the third codon position in coding sequences (GC3) and the mean intron GC content are positively correlated, though again, not significantly. Finally, we find evidence for natural selection or biased gene conversion favouring AT to GC using the maximum-likelihood method that takes the recent population expansion into account. This suggests that the signal showing GC favoured might be partly obscured by the distortions in allele frequencies due to the population expansion. In any case, if would not be surprising if GC mutations were favoured, as several studies in *Drosophila* have shown selection or biased gene conversion favouring GC variants (Marais et al. 2003; Casillas et al. 2007; Haddrill and Charlesworth 2008).

The estimate of natural selection or biased gene conversion favouring AT to GC polymorphisms in introns (γ=0.35) is smaller than that for natural selection favouring U to P polymorphisms (γ=1.51), suggesting that it is weaker than selection for codon usage bias and more likely to be affected to population size changes. Zeng and Charlesworth (2009) investigated the effects of a population expansion on patterns of polymorphism. They found that, just after a population expansion, the chance of mutations of the unfavoured allele a to the favoured allele A is higher than at equilibrium with the new population size. This, combined with the stronger selection for codon usage bias, could explain why we observe similar numbers of AT > GC and GC > AT polymorphisms,

although we expect more GC > AT polymorphisms (due to mutational bias), and more P > U than U > P polymorphisms at equilibrium. Indeed, the stronger selection for preferred codons vs. weaker selection for GC variants might explain why codon usage bias appears to have recovered more quickly from the effects of the population expansion, in terms of numbers of polymorphisms. The simulations from Zeng and Charlesworth (2009) suggest that the γ ($2N_e s$) estimates (for both selection for codon usage bias and selection for GC variants) will increase shortly after the population expansion and then, as time progresses from the population expansion, decline to the new equilibrium value.

5.4.3 Recombination

Our results do not show a significant difference in recombination rate between introns and exons. This suggests that, if the same forces act on introns and exons, the same patterns should be observed. The finding that there is a bias toward GC variants in both coding and non-coding sequences suggests that biased gene conversion may be acting on introns, and also on coding sequences, in conjunction with selection for codon usage bias.

6 Gene expression on the fourth chromosome versus other chromosomes in 7 *Drosophila* species.

Contributing authors:

- I collected the data and performed the analyses.
- A. Betancourt helped with the data collection and analysis.
- B. Charlesworth advised on the project.

6.1 Introduction

The fourth chromosome in *Drosophila melanogaster* is very small, and is therefore often called the dot chromosome (Ashburner 1989); it corresponds to Muller element F. Despite their separate evolutionary histories, many *Drosophila* species have maintained an equivalent of the *D. melanogaster* dot chromosome. In the majority of *Drosophila* species, the Muller element F is similar to that of *D. melanogaster*, maintained as a small dot chromosome. The only exception in *Drosophila* is *D. ananassae*, where the Muller F is a larger chromosome with two distinct arms (Kikkawa 1938; Schaeffer et al. 2008). The Muller element F in *D. willistoni* has fused with Muller element C (Papaceit and Juan 1998; Schaeffer et al. 2008).

The Muller element F is completely lacking crossing-over in *D. melanogaster* (Ashburner et al. 2005), and it is generally assumed that it has low crossing-over rate in other *Drosophila* species as well (Riddle and Elgin 2006). Low crossing-over is known to affect patterns of molecular evolution due to Hill-Robertson interference (Hill and Robertson, 1966; Felsenstein, 1974; Maynard Smith and Haigh 1974). In particular, low crossing-over is associated with high non-synonymous divergence, low nucleotide site diversity, and a lower efficacy of natural selection, both purifying and adaptive (Betancourt and Presgraves 2002; Haddrill et al. 2007; Betancourt et al. 2009). As mentioned in Chapter 5, recombination can also affect biased gene conversion, a non-

selective process.

Gene expression is positively correlated with codon usage bias (Duret and Mouchiroud 1999; Marais et al. 2001, 2004) and negatively correlated with levels of non-synonymous divergence d_N (Marais et al. 2004; Larracuente et al. 2008). Therefore, if genes on the Muller element F have lower gene expression than other chromosomes, we would expect to observe low codon usage bias and high non-synonymous divergence on this chromosome. Haddrill et al. (2008b) suggested that, if genes in non-crossover regions such as the Muller F element have lower gene expression, the correlations of codon usage bias and non-synonymous divergence with gene expression might explain the results of low codon usage and elevated rates of non-synonymous divergence in non-crossover regions (Haddrill et al. 2008b).

The level of gene expression is of major significance in the evolution of DNA sequences and it can be measured in several ways. Expressed Sequence Tags (EST) and microarrays are two widely used methods. An EST (Adams et al. 1991) is a short sub-sequence of a transcribed cDNA sequence, so it represents a portion of an expressed gene, and the EST counts give estimates of gene expression. DNA microarrays (Pease et al. 1994; Brown et al. 1999; Duggan et al. 1999) are more recent and consist of an arrayed series of thousands of DNA oligonucleotides or probes that are used to hybridize a cDNA sample. Hybridization is detected and quantified using fluorescent targets.

Haddrill et al. (2008b) showed that gene expression (as measured by EST counts) is actually higher for genes on the 4th chromosome than for genes on other chromosomes. This might be explained by the lack of regulation of gene expression in low-recombination regions or by compensation for reduced protein function (Haddrill et al. 2008b). I attempted to replicate these results for all sequenced *Drosophila* species, using microarray data but I found the opposite result: lower expression for 4th chromosome genes than for other genes. This result was highly significant and consistent across all species.

6.2 Methods

We retrieved gene expression data for *D. simulans*, *D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. mojavensis* and *D. virilis* from the GEO database. This database was generated by Zhang et al. (2007), who recently performed microarray experiments to investigate sex-biased expression of orthologues and species-restricted genes in *Drosophila* (data available at NCBI GEO database (Edgar et al. 2002), accession GSE6640). We used the log2 transformed signal intensities after VSN normalization. VSN stands for variance stabilization and normalization (Huber et al. 2002) and is necessary before analyzing microarray data. These data were available for different numbers of males and females, so we calculated the weighted average of these values for each gene. Table 6.1 shows the number of genes used in the final dataset for each species in the dot chromosome, and on other chromosomes.

We also extracted EST counts from Unigene, to check how the two measures for gene expression compare. We obtained the file for all EST counts in *D. melanogaster* on the ftp server of Unigene, and extracted the gene identification and EST counts for all genes. We also extracted EST counts from a single library that used 454 sequencing (nebulized cDNA library) on an adult female fly (Lib. 21597).

To assign each gene to the appropriate Muller element, we first used the pre-computed tables from FlyBase to assign each gene to its scaffold in the genome assembly, and then we used data from Schaeffer et al. (2008) to assign each scaffold to a Muller element.

We then tested whether there was a difference in gene expression level between genes on Muller element F and genes on other Muller elements. All analyses were performed in R (http://cran.r-project.org/).

6.3 Results

6.3.1 Microarray data

The mean expression was significantly lower on the Muller element F than on other Muller elements for each of the seven *Drosophila* species studied (Wilcoxon tests, $p<0.029$) (Table 6.1 and Figure 6.1).

Table 6.1: Number of genes used for each species on the 4th chromosome and on other Muller elements.

Species	n (F element)	n (other)	Mean expr. F (SD)	Mean expr. other (SD)	p-value (Wilcoxon test)
D. melanogaster	73	12,389	9.20 (1.26)	9.70 (1.00)	3.84×10^{-5}
D. simulans	49	11,842	8.12 (1.02)	8.35 (0.93)	0.029
D. yakuba	66	12,867	8.64 (1.33)	9.24 (1.03)	2.26×10^{-6}
D. pseudoobscura	69	18,405	8.87 (1.19)	9.22 (0.85)	5.24×10^{-4}
D. virilis	75	13,722	9.89 (1.10)	10.31 (0.96)	2.79×10^{-6}
D. mojavensis	72	11,445	8.98 (1.25)	9.36 (0.95)	4.75×10^{-4}
D. ananassae	68	12,486	8.75 (1.57)	9.16 (1.07)	3.14×10^{-3}

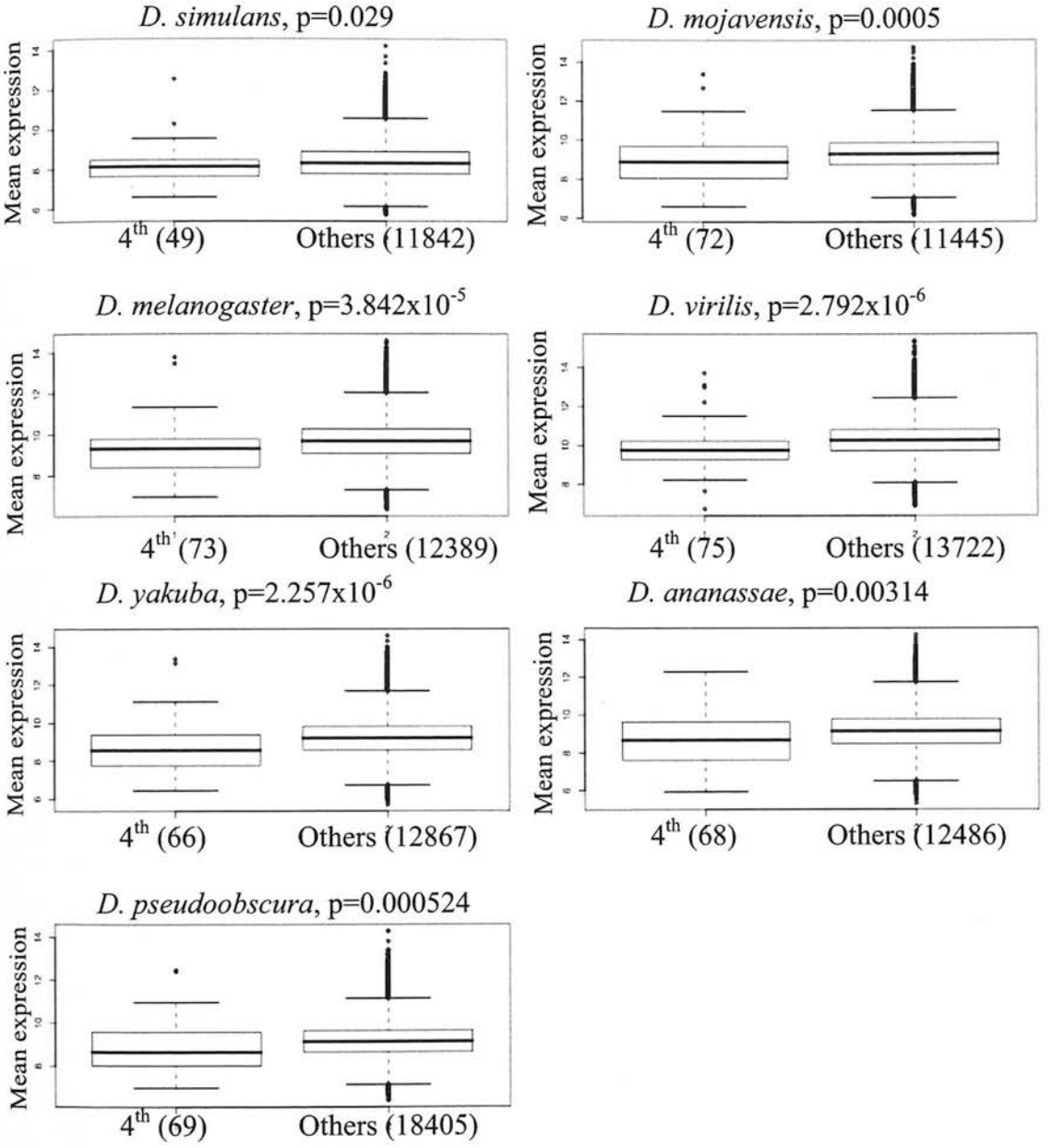

Figure 6.1: Boxplots for the microarray expression data for 7 species of *Drosophila*, comparing the values for the 4th chromosome (Muller element F) to other chromosomes.

6.3.2 Microarray vs. EST counts for *D. melanogaster*

Since the expression level using microarray data was lower on the Muller element F than on other Muller elements, in contrast to the results of Haddrill et al. (2008b), we investigated the same genes using EST counts to see whether the difference was due to the set of genes. We obtained EST counts for 20 genes on Muller element F and for 1770

genes on other Muller elements.

A factor that could bias the results is the gender of the flies used and the tissue where gene expression was measured. Therefore we chose to compare the mean expression data for whole females with EST counts from a whole female.

Using EST counts from the Library 21597, we obtained the same result as in Haddrill et al. (2008b), the level of gene expression was significantly higher on Muller element F than on other Muller elements (Wilcoxon test, $p= 2.9x10^{-4}$) (Figure 6.2). When we looked at the difference in microarray expression between Muller element and other Muller elements for the subset of genes with EST counts from the Library 21597, we find the same difference as with the whole dataset: microarray expression is lower on the Muller element F than on other Muller elements (Wilcoxon test, $p=2.4x10^{-5}$). We looked at the correlation between this subset of EST data and the female gene expression from GEO; they are significantly positively correlated both overall (Pearson r = 0.52, $p< 2.2x10^{-16}$) (Figure 6.3) and within the 4th and non-4th chromosome categories considered separately (4th: Pearson r = 0.45, p =0.049; non-4th: Pearson r =0.52, $p< 2.2x10^{-16}$).

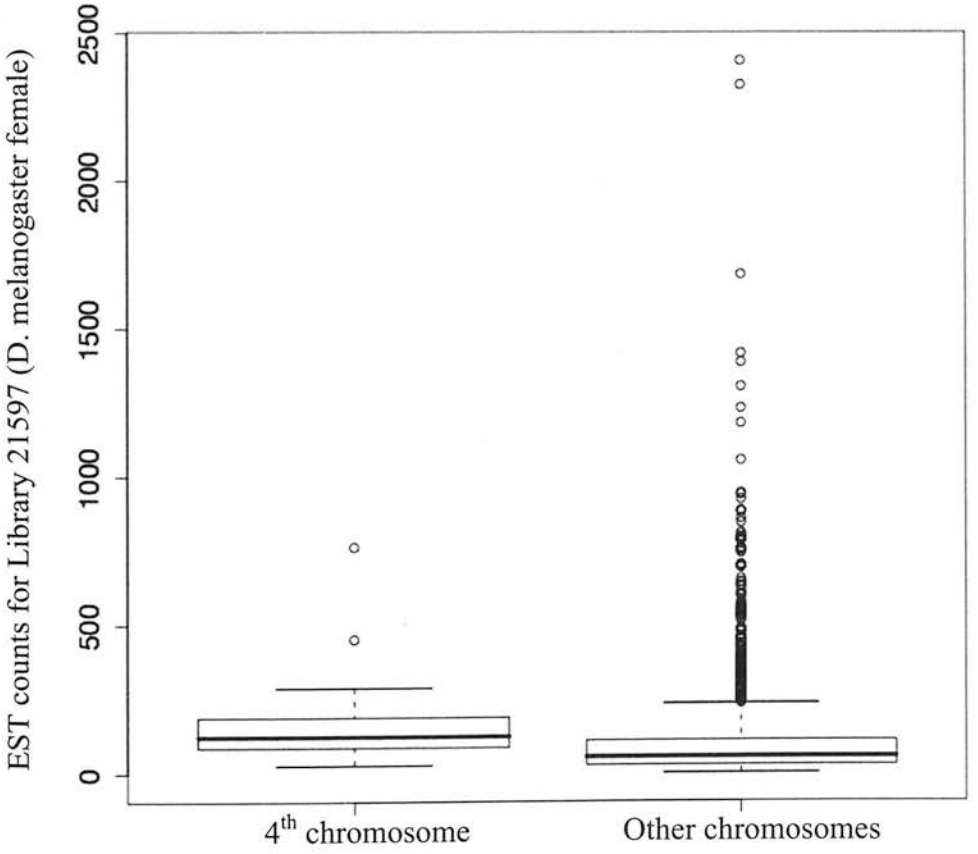

Figure 6.2: Boxplot for the EST counts in a whole *D. melanogaster* female (library 21597) comparing the values for the 4th chromosome (Muller element F) to other chromosomes.

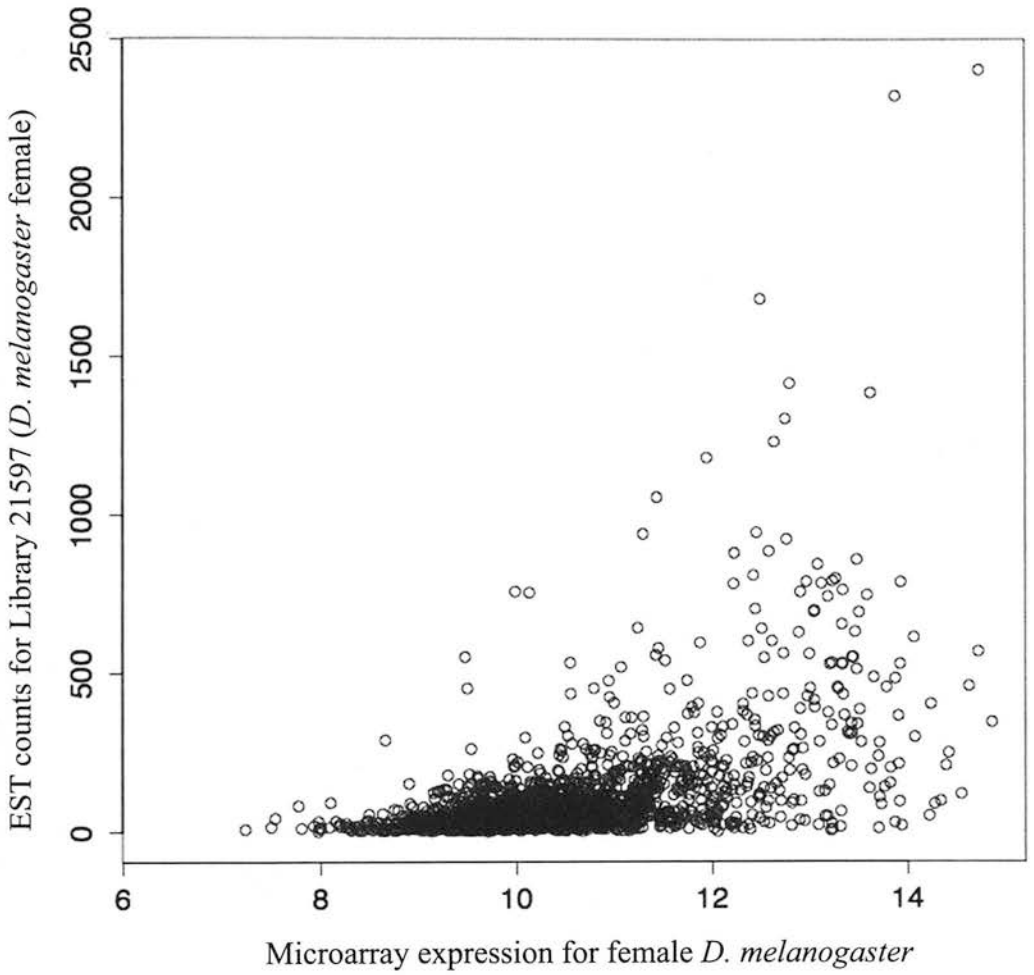

Figure 6.3: Plot of EST counts for library 21597 (whole *D. melanogaster* female) against microarray expression data for *D. melanogaster* females. The positive correlation between these two datasets is highly significant (Spearman's rank correlation, rho=0.53, $p<2.2x10^{-16}$).

6.3.3 Gene length effect

Munoz et al. (2004) found that long genes tend to be over-represented in EST datasets compared with their expression as measured with microarrays. If genes in regions of low recombination are longer, that could explain the discrepancy. Munoz et al. (2004) assume that microarrays are more reliable, and they assign the bias to the EST dataset. Therefore,

simply controlling for gene length could remove the bias. As the median length was 1227 bp, we separated the dataset into 6224 long (1227 bp and over, including 55 genes on the 4th chromosome and 6169 on other chromosomes) and 6218 short genes (less than 1227 bp, including 18 genes on the 4th chromosome and 6200 on other chromosomes).

The difference in coding sequence length between *D. melanogaster* fourth chromosome genes and non-fourth chromosome genes was significant, with longer genes found on the fourth chromosome (Wilcoxon test, W = 613934.5, p-value = 1.10×10^{-7}) (Table 6.2).

We investigated the effect of gene length by looking at the difference in gene expression between 4th and other chromosomes in short and long genes separately. Using microarray data, the difference was still in the same direction (lower expression on the 4th) in both categories (Tables 6.3) but only significant in long genes (Wilcoxon test, $p=6.86 \times 10^{-5}$). Using EST data, the difference was still opposite (higher expression on the 4th) to the microarray data (Table 6.4) for all genes and long genes, but only significant in long genes (Wilcoxon test, $p= 1.97 \times 10^{-5}$). The difference is in the opposite direction for short genes, but not significant. Gene length is positively correlated with EST counts (Spearman r_S = 0.128, $p=5.9 \times 10^{-8}$) but negatively correlated with microarray data (Spearman r_S = -0.081, $p<2.2 \times 10^{-16}$). Since genes are significantly longer on the 4th chromosome, we tried to determine whether the difference in EST counts between 4th and other chromosomes was solely due to gene length. We calculated the partial correlation coefficient between EST counts and gene length after accounting for the chromosome (using 2 for 4th chromosome, 1 for other chromosomes). The partial correlation is quite weak, but positive and significant (Kendall τ=0.066, $p=2.81 \times 10^{-5}$), suggesting that the difference is not only due to the gene length effect.

Table 6.2: Mean gene length and standard error for *D. melanogaster* genes used in this study. Long genes are 1227bp and above, short genes are less than 1227bp.

Category	4th chromosome	Other
All genes	2875.4 (399)	1615.3 (15.3)
Long genes	3588.8 (493)	2510.4 (25.9)
Short genes	695.8 (66.3)	724.7 (3.6)

Table 6.3: Microarray expression data and standard error for Library 21597 (*D. melanogaster* female). Long genes are 1227bp and above, short genes are less than 1227bp.

Category	4th chromosome	Other
All genes	9.06 (0.16)	9.60 (0.01)
Long genes	8.93 (0.16)	9.52 (0.01)
Short genes	9.46 (0.42)	9.69 (0.02)

Table 6.4: Mean number of EST counts and standard error for Library 21597 (*D. melanogaster* female). Long genes are 1227bp and above, short genes are less than 1227bp.

Category	4th chromosome	Other
All genes	176 (37.9)	107.1 (4)
Long genes	216.2 (53)	100.7 (5.4)
Short genes	101.4 (33.3)	113.6 (6)

6.4 Discussion

In this Chapter, we attempted to replicate the results from Haddrill et al. (2008b) who found that genes on the 4th chromosome had higher expression levels as measured by EST counts. We used microarray data from Zhang et al. (2007) for seven *Drosophila* species. Our results show that, in contrast to the finding by Haddrill et al. (2008b), genes on the 4th chromosome had lower gene expression than genes on other chromosomes. These contradictory results were unexpected, and suggest that there may be a bias in EST counts data for gene expression. To test whether the difference was due to a specific dataset, we used microarray data for *D. melanogaster* females only and compared it with an EST library for one whole *D. melanogaster* female to correct for potential gender and tissue differences. The results using these specific datasets remained the same, suggesting that the nature of the gene expression measure was the source of the difference.

Munoz et al. (2004) proposed that EST datasets are biased towards long genes, so that gene length should be corrected for when using EST data in gene expression analyses. To correct for gene length, we separated the genes into short and long genes and calculated partial correlation coefficients to account for gene length, but the results suggest that the difference in EST counts between 4th and other chromosomes are not

solely due to the effect of gene length.

Since Munoz et al. (2004) proposed that there is a bias in EST datasets, it is necessary to evaluate the microarray results. A trend to over-expression of genes on the 4th chromosome similar to that in Haddrill et al. (2008b) has been found by John H. Malone (NIH, personal comm.) using both microarray data from Zhang et al. (2007) and RNA sequence data for heads from three *Drosophila* species. However, we found a consistently lower gene expression level on the 4th chromosome than on other chromosomes across seven *Drosophila* species. This difference could be attributed to the main characteristic of the 4th chromosome, the lack of crossing over. Recombination has been shown to reduce the efficiency of natural selection, so that selection for codon usage bias is lower in low recombination regions due to Hill-Robertson effects (Charlesworth 1994; Akashi et al. 1998; Kreitman and Comeron 1999; Hey 1999). It is also well documented that codon usage bias is usually higher in highly expressed genes (Gouy and Gautier 1982; Sharp and Li 1986; Duret and Mouchiroud 1999; Marais et al. 2001; 2004). Haddrill et al. (2008b) did not find a significant difference in gene expression level between different categories of recombination rates, and the only significant difference, despite its opposite direction, was observed between non-crossover genes, especially on the 4th chromosome, and other genes. This suggests that only the 4th chromosome has been lacking crossing-over for long enough to have noticeable effects on evolutionary patterns. Studies on yeast have shown that expression levels are positively correlated with meiotic DSB rates (Pál et al. 2001), and with crossover rates (Weber and Hurst 2010), suggesting that there is some mechanism linking these two variables.

We also need to consider the possibility that probes in GC-rich genes may bind microarrays more strongly (Royce et al. 2007). Since longer genes have lower codon usage bias and hence lower GC content, this could account for the conflict between EST and microarray data. Indeed, longer genes on the 4th chromosome will have a lower GC content and therefore reduced levels of expression using microarrays.

7 Conclusions

7.1 Summary

The evolution of non-coding DNA sequences has only recently been the focus of population genetics studies, and an interesting finding is that non-coding sites appear to be under selective constraint (Bergman and Kreitman 2001; Andolfatto 2005; Halligan and Keightley 2006), with levels of divergence close to those in synonymous sites. This suggests that a large part of eukaryotic genomes are functional. In this thesis, we investigated various factors affecting molecular evolution, particularly in non-coding sequences.

In Chapter 2, we obtained sequences for a large number of genes in *D. miranda* from BAC sequences, and compared these with sequences from its close relative, *D. pseudoobscura*. As in previous studies in *D. melanogaster* (Parsch 2003; Marais et al. 2005; Haddrill et al. 2005; Bachtrog and Andolfatto 2006; Halligan and Keightley 2006), we found a negative relationship between intron length and intron divergence, suggesting that longer introns are under selective constraint. Surprisingly, when investigating the effect of intron length on divergence and polymorphism in a smaller dataset in *D. americana* in Chapter 3, we found that short introns actually show lower levels of polymorphism and divergence than long introns, suggesting higher constraint on short introns in this particular species.

An interesting finding in *Drosophila* is that a non-neutral process affects both coding and non-coding sequences and favours AT to GC polymorphisms. It is not possible to distinguish between selection for GC variants or biased gene conversion that repairs double-stranded breaks preferentially with G or C nucleotides (Marais 2003). In Chapter 5, we show that this non-neutral force is acting on introns in *D. americana*. This further suggests that introns are not evolving neutrally in *D. americana*.

In *Drosophila*, a polymorphism deletion bias has been well documented, leading to the idea of selection favouring insertions to compensate for the DNA loss (Parsch

2003). We investigated patterns of indel polymorphisms in Chapter 4 and did not find evidence for natural selection on either insertions or deletions, although insertions show slightly higher frequencies than deletions.

There is strong evidence for selection for codon usage bias in *Drosophila melanogaster* (Shields et al. 1988; Akashi 1995). As found in Maside and Charlesworth (2004), we confirm that there is selection for codon usage bias in *D. americana* in Chapter 5. Since we showed in Chapter 3 that introns are subject to similar evolutionary forces as synonymous sites, this further suggests that introns are selectively constrained.

Recombination can affect non-neutral processes. Higher recombination rates increase the efficiency of natural selection as well as the rate of biased gene conversion. Differences in recombination rates between coding and non-coding sequences could therefore affect patterns of evolution. However, we show in Chapter 5 that recombination rate estimates are similar in introns and exons in *D. americana*, suggesting that, if the same forces act on introns and exons, the same patterns should be observed.

In Chapter 6, we investigated the levels of gene expression in seven *Drosophila* species between the 4th chromosome (Muller element F), which completely lacks crossing-over, and other chromosomes. Gene expression is usually associated with stronger selection for codon usage bias. Our results using microarray data suggest that gene expression is lower on the 4th chromosome, as opposed to the finding of Haddrill et al. (2008b), who used EST counts as a measure of gene expression. These opposite results suggest a bias in gene expression datasets. A lower expression for genes on the 4th chromosome could explain the correlations between codon usage bias and recombination and gene expression.

We also found a negative correlation between the rate of non-synonymous substitutions and codon usage bias in *D. miranda* in Chapter 2, suggesting that fast-evolving genes have a lower codon usage bias, consistent with strong positive selection interfering with weak selection for codon usage. This correlation has been documented in *D. melanogaster* (Betancourt and Presgraves 2002; Marais et al. 2004; Bierne and Eyre-Walker 2006; Andolfatto 2007; Bachtrog 2008).

The demographic history of a population can affect nucleotide sequences and hence interfere with signals of natural selection (Hahn et al. 2002). In Chapter 4, we showed that not taking into account a change in population size can drastically alter the conclusions of tests for natural selection, even when the population studied is thought to have been demographically stable for a long time, as was the case here with *D. americana*.

7.2 Future directions

The conclusions from Chapter 2 could gain power with a whole genome sequence for *Drosophila miranda*. With the advances in high-throughput sequencing technologies, this could be available in the near future and allow whole-genome comparisons to determine whether non-significant correlations are genuine but too weak to be detected in our dataset.

An annotated genome sequence for *D. miranda* would also allow to accurately determine how much constraint in intergenic sequences is due to UTRs. It should also be easier to detect a potential effect of intron position on divergence, and test the hypothesis that first introns are more selectively constrained than introns at the end of coding sequences. Extra sequence data should also make it possible to study functional constraint at greater distances from coding sequences.

The main issue when using polymorphism data is that the production of relatively small datasets by classic sequencing methods takes a long time. The publication of the 12 *Drosophila* genomes (Clark et al. 2007) has provided much information about evolutionary processes on a large scale and for a large range of phylogenetic distances. With high-throughput sequencing methods, these genome sequences could soon be supplemented by sequences for more strains in various populations of *Drosophila*, with the potential for much more powerful studies of polymorphism and divergence.

Such methods could also be applied to obtain sequences for various chromosomal regions. For example, most studies investigate genes and non-coding sequences in

euchromatin. However, it has been shown that heterochromatin, which constitutes centromeres and telomeres, has been associated with several functions, from gene regulation to the protection of the integrity of chromosomes (Grewal and Jia 2007); some of these roles can be attributed to the dense packing of DNA, which makes it less accessible to protein factors that usually bind DNA or its associated factors. Such functions should involve selective constraint and therefore leave non-neutral patterns of polymorphism and divergence on DNA sequences. Heterochromatin could therefore provide a new class of non-coding sequence subject to natural selection to a certain extent.

Bibliography

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF et al. (1991) Complementary DNA sequencing : expressed sequence tags and human genome project. *Science* **252**(5013) :1651–1656

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amantides PG, et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287**:2185–2195

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**:927–935

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics* **139**:1067–1076

Akashi H (1999) Inferring the fitness effects of DNA mutations from polymorphism and divergence data: Statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**:221–238

Akashi H and SW Schaeffer (1997) Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. *Genetics* **146**:295–307

Akashi H and A Eyre-Walker (1998) Translational selection and molecular evolution *Curr Op Genet Dev* **8**(6):688–693

Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**:1149–1152

Andolfatto P (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res* **17**:1755–1762

Andolfatto P and M Przeworski (2000) A genome-wide departure from the standard neutral model in natural populations of Drosophila. *Genetics* **156**:257–268

Arnone MI and EH Davidson (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**:1851–1864

Ashburner M (1989) *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Ashburner M, KG Golic and RS Hawley (2005) *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Babbitt GA and Y Kim (2008) Inferring natural selection on fine-scale chromatin organization in yeast. *Mol Biol Evol* **25**(8):1714–1727

Bachtrog D (2008) Similar rates of protein adaptation in *Drosophila miranda* and *D. melanogaster,* two species with different current effective population sizes. *BMC Evol Biol* **8**:334

Bachtrog D and P Andolfatto (2006) Selection, recombination and demographic history in *Drosophila miranda. Genetics* **174**:2045–2059

Bachtrog D, Hom E, Wong KM, Maside X, de Jong P (2008) Genomic degradation of a young Y chromosome in *Drosophila miranda. Genome Biol* **9**:R30

Barrio E, Latorre A, Moya A and FJ Ayala (1992) Phylogenetic reconstruction of the *Drosophila obscura* group, on the basis of mitochondrial DNA. *Mol Biol Evol* **9**:621–635

Bartolomé C, Maside X, Yi S, Grant AL and B Charlesworth (2005) Patterns of selection on synonymous and nonsynonymous variants in *Drosophila miranda. Genetics* **169**:1495–1507

Bartolomé C and B Charlesworth (2006a) Rates and patterns of chromosomal evolution in *Drosophila pseudoobscura* and *D. miranda. Genetics* **173**:779–791

Bartolomé C and B Charlesworth (2006b) Evolution of amino-acid sequences and codon usage on the *Drosophila miranda* neo-sex chromosomes. *Genetics* **174**:2033–2044

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, Pachter L, Myers E, Langley CH (2007) Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans. PLoS Biol* **5**:e310

Bergman CM and M Kreitman (2001) Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res* **11**:1335–1345

Bergman CM, Pfeiffer BD, Rincón-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, Stapleton M, Wan K, George RA, de

Jong PJ, Botas J, Rubin GM and SE Celniker (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol* **3**(12):research0086.1–0086.20

Betancourt AJ and DC Presgraves (2002) Linkage limits the power of natural selection in *Drosophila*. *Proc Natl Acad Sci USA* **99**(21):13616–13620

Betancourt AJ, Welch JJ and B Charlesworth (2009) Reduced effectiveness of selection caused by a lack of recombination. *Curr Biol* **19**:655–660

Bierne N and A Eyre-Walker (2003) The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: Implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* **165**:1587–1597

Bierne N and A Eyre-Walker (2006) Variation in synonymous codon use and DNA polymorphism within the *Drosophila* genome. *J Evol Biol* **19**:1–11

Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature* **447**:799–816

Bradnam KR and I Korf (2008) Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE* **3**(8):e3093

Bray N and L Pachter (2004) MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res* **14**:693–699

Brown PO and D Botstein (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genet* **21**:33–37

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**:897–9070

Carvalho AB and AG Clark (1999) Intron size and natural selection. *Nature* **401**:344.

Casillas S, Barbadilla A and CM Bergman (2007) Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Mol Biol Evol* **24**(10):2222–2234

Chamary JV and LD Hurst (2004) Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol* **21**:1014–1023

Charlesworth B (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res* **63**:213–227

Charlesworth B (1996) The changing size of genes. *Nature* **384**:315–316.

Charlesworth B (2001) The effect of life-history and mode of inheritance on neutral genetic variability. *Genet Res Camb* 77:153–166.

Charlesworth B, Coyne JA and NH Barton (1987) The relative rates of evolution of sex chromosomes and autosomes. *Am Nat* **130**(1):113–146

Charlesworth B, Morgan MT and D Charlesworth (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**:1289–1303

Charlesworth D, Charlesworth B and MT Morgan (1995) The pattern of neutral molecular variation under the background selection model. *Genetics* **141**:1619–1632

Chen Y and W Stephan (2003) Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene. *Proc Natl Acad Sci USA* **100**(20):11499–11504

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. (416 co-authors) (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**:203–218

Comeron JM (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J Mol Evol* **41**:1152–1159

Comeron JM (2001) What controls the length of noncoding DNA? *Current Opinion in Genetics & Development* **11**:652–659.

Comeron JM and M Aguadé (1996) Synonymous substitutions in the *Xdh* gene of Drosophila: Heterogeneous distribution along the coding region. *Genetics* **144**:1053–1062

Comeron JM, Kreitman M and Aguadé (1999) Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. *Genetics* **151**:239–249

Comeron, JM and M Kreitman (2000) The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. *Genetics* **156**: 1175–1190.

Comeron JM, Williford A and RM Kliman (2008) The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity* **100**:19–31

Cutter AD and B Charlesworth (2006) Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. *Curr Biol* **16**:2053–2057

Dean MD and JWO Ballard (2004) Linking phylogenetics with population genetics to reconstruct the geographic origin of a species. *Mol Phylogen Evol* **32**:998–1009

Dermitzakis ET, Bergman CM and AG Clark (2003) Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. *Mol. Biol. Evol.* **20**(5): 703–714

Díaz-Castillo C and KG Golic (2007) Evolution of gene sequence in response to chromosomal location. *Genetics* **177**:359–374

Dos Reis M and L Wernisch (2009) Estimating translational selection in eukaryotic genomes. *Mol Biol Evol* **26**(2):451–461

Duret L (2001) Why do genes have introns? Recombination might add a new piece to the puzzle. *Trends Genet* **17**(4):172–175

Duret L and D Mouchiroud (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Cænorhabditis*, *Drosophila*, and *Arabidopsis*. Proc *Natl Acad Sci USA* **96**(8):4482–4487

Drummond DA and CO Wilke (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**:341–352

Duggan DJ, Bittner M, Chen Y, Meltzer and JM Trent (1999) Expression profiling using cDNA microarrays. *Nature Genet* **21**:10–14

Emberly E, Rajewsky N and ED Siggia (2003) Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics* **4**:57

Edgar R, Domrachev M and AE Lash (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**(1):207–210

The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.

Eyre-Walker A (1996) Synonymous codon bias related to gene length in *Escherichia coli*: Selection for translational accuracy? *Mol Biol Evol* **13**(6):864–872

Eyre-Walker A (2006) The genomic rate of adaptive evolution. Trends in Ecology and Evolution **21**(10):569–575

Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics* **78**(2):737–756

FLYBASE: A database of the Drosophila genome [http://www.flybase.org]

Fu XY and WH Li (1993) Statistical tests of neutrality of mutations. *Genetics* **133**:693–709

Galtier N, Gouy M and C Gautier (1996) SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Comput. Applic. Biosci.* **12**: 543–548

Galtier N, Piganeau G, Mouchiroud D and L Duret (2001) GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics* **159**:907–911

Galtier N, Bazin E and N Bierne (2006) GC-biased segregation of noncoding polymorphisms in *Drosophila*. *Genetics* **172**: 221–228

Galtier N and L Duret (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet* **23**(6):273–277

Galtier N, Duret L, Glémin S and V Ranwez (2009) GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Cell* **25**(1):1–5

Gillespie JH (1991) The causes of molecular evolution Oxford University Press, New York

Gouy M and C Gautier (1982) Codon usage bias in bacteria: correlation with gene expressivity. *Nucl Ac Res* **10**(22):7055–7074

Gregory TR (2004) Insertion-deletion biases and the evolution of genome size. *Gene* **324**:15–34.

Grewal SIS and S Jia (2007) Heterochromatin revisited. *Nature Rev Genet* **8**:35–46

Gutz H and JF Leslie (1976) Gene conversion: A hitherto overlooked parameter in population genetics. *Genetics* **83**:861–866

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Charlesworth B, and PD Keightley (2007) Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445**:82–85

Haddrill PR, Thornton KR, Charlesworth B and P Andolfatto (2005) Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol* **6**:R67

Haddrill PR, Halligan DL, Tomaras D and B Charlesworth (2007) Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. *Genome Biol* **8**:R18.

Haddrill PR and B Charlesworth (2008) Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. *Biol Lett* **4**:438–441

Haddrill PR, Bachtrog D, and P Andolfatto (2008a) Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Mol Biol Evol* **25**(9):1825–1834

Haddrill PR, Waldron FM and B Charlesworth (2008b) Elevated levels of expression associated with regions of the *Drosophila* genome that lack crossing over. *Biol Lett* **4**:758–761.

Hahn MW, Rausher MD and CW Cunningham (2002) Distinguishing between selection and population expansion in an experimental lineage of bacteriophage T7. *Genetics* **161**:11–20

Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* **210**:1518–1525

Halligan DL, Eyre-Walker A, Andolfatto P and PD Keightley (2004) Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. *Genome Res* **14**:273–279

Halligan DL and PD Keightley (2006) Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res* **16**:875–884

Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16(9):369–372

Hartl DL, Moriyama EN and SA Sawyer (1994) Selection intensity for codon bias. *Genetics* **138**:227–234

Heger A and CP Ponting (2007) Variable strength of translational selection among 12 Drosophila species. *Genetics* **177**:1337–1348

Hey J (1999) The neutralist, the fly and the selectionist. *Trends Ecol Evol* **14**(1):35–38

Hill WG and A Robertson (1966) The effect of linkage on limits to artificial selection. *Genet Res* **8**:269–294

Hsu TC (1952) Chromosomal variation and evolution in the virilis group of Drosophila. Univ. Texas Pub. 5204:35–72

Huber W, von Heydebreck A, Sültmann H, Poutska A and M Vingron (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**(Suppl 1):S96–S104

Hudson RR (1991) Gene genealogies and the coalescent process. In Oxford Surveys in Evolutionary Biology (Eds D. Futuyama and J. Antonovics), volume 7, pp. 1-44. Oxford University Press.

Hudson RR, Kreitman M and M. Aguade (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**:153–159

Ikemura T (1981) Correlation between the abundance of *E. coli* transfer RNAs and the occurrence of the respective codon in the protein genes. *J Mol Biol* **146**:1–21

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**(1):13–34

Jukes TH and CR Cantor (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp. 21–132

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J and E Segal (2009) The DNA-encoded nucleosome

organization of a eukaryotic genome. *Nature* **458**:362–366

Keightley PD and DJ Gaffney (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc Natl Acad Sci USA* **100**(23):13402–13406

Keightley PD and T Johnson (2004) MCALIGN: stochastic alignment of noncoding DNA sequences based on an evolutionary model of sequence evolution. *Genome Res* **14:**442–450

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S and ML Blaxter (2009) Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* **19**(7):1195–1201

Kikkawa H (1938) Studies on the genetics and cytology of *Drosophila ananassae*. *Genetica* **20**(5-6):458–516

Kimura M (1968) Evolutionary rate at the molecular level. *Nature* **217**(5129):624–626

Kimura M (1983) The neutral theory of molecular evolution. Cambridge, U.K.: Cambridge Univ. Press.

Kirby DA, Muse SV and W Stephan (1995) Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci USA* **92**(20):9047–9051

Kliman RM and J Hey (1994) The effects of mutation and natural selection on codon usage bias in the genes of Drosophila. *Genetics* **137**:1049–1056

Kreitman M and J Comeron (1999) Coding sequence evolution. *Curr Op Genet Dev* **9**:637–641

Kurland CG (1987) Strategies for efficiency and accuracy in gene expression. *Trends Biochem Sci* **12**(4):126–128

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B and AG Clark (2008) Evolution of protein-coding genes in *Drosophila*. *Trends Genet* **24**(3):114–123

Leicht BG, Muse SV, Hanczyc M and AG Clark (1995) Constraints on intron evolution in the gene encoding the myosin alkali light chain in Drosophila. *Genetics* **139**:299–308

Li WH (1987) Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons. *J Mol Evol* **24**(4):337–345

Ludwig MZ, Bergman C, Patel NH and M Kreitman (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**:564–567

Lynch M (2007) *The Origins of Genome Architecture* (Sinauer, Sunderland, MA)

Lynch (2010) Rate, molecular spectrum and consequences of human mutations. *Proc Natl Acad Sci USA* **107**(3):961–968

Majewski J and J Ott (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res* **12**:1827–1836

Mantel N and W. Haenszel (1959) Statistical aspects of the analysis of data from the retrospective analysis of disease. *J Natl Cancer Inst* **22**:719

Marais G (2003) Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics* **19**:330–338

Marais G, Mouchiroud D and L Duret (2001) Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc Natl Acad Sci USA* **98**(10):5688–5692

Marais G, Mouchiroud D and L Duret (2003) Neutral effect of recombination on base composition in *Drosophila. Genet Res* **81**:79–87

Marais G, Domazet-Lošo T, Tautz D and B Charlesworth (2004) Correlated evolution of synonymous and nonsynonymous sites in *Drosophila. J Mol Evol* **59**:771–779

Marais G, Nouvellet P, Keightley PD and B Charlesworth (2005) Intron size and exon evolution in *Drosophila. Genetics* **170**:481–485

Marion de Procé S, Halligan DL, Keightley PD and B Charlesworth (2009) Patterns of DNA-sequence divergence between *Drosophila miranda* and *D. pseudoobscura. J Mol Evol* **69**(6):601–611

Maroni G (1994) The organization of Drosophila genes. *DNA Seq* **4**(6):347–354

Maruyama T and PA Fuerst (1984) Population bottlenecks and non-equilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. *Genetics* **108**:745–763

Maruyama T and PA Fuerst (1985) Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics* **111**:675–689

Maside X, AW Lee and B Charlesworth. (2004) Selection on codon usage in *Drosophila americana. Curr Biol* **14**:150–154

Maside X and B Charlesworth (2007) Patterns of molecular variation and evolution in *Drosophila americana* and its relatives. *Genetics* **176**:2293–2305

Maynard-Smith J and J Haigh (1974) Hitch-hiking effect of a favorable gene. *Genet Res* **23**:23–35

McAllister BF (2002) Chromosomal and allelic variation in *Drosophila americana*: selective maintenance of chromosomal cline. *Genome* **45**(1):13–21

McAllister BF and B Charlesworth (1999) Reduced sequence variability on the *Neo-Y* chromosome of *Drosophila americana americana. Genetics* **153**:221–233

McAllister BF and AL Evans (2006) Increased nucleotide diversity with transient Y linkage in *Drosophila americana. PLoS One* **1**(1):112

McDonald JH and M Kreitman (1991) Adaptive evolution at the *Adh* locus in *Drosophila. Nature* **351**:652–654

McVean GAT and J Vieira (2001) Inferring patterns of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. *Genetics* **157**:245–257

McVean GAT, Awadalla P and P Fearnhead (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**:1231–1241

Mount SM, Burks C, Hertz G, Stormo GD, White O and C Fields (1992) Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucl Ac Res* **20**(16):4255–4262

Moriyama EN, Powell JR (1998) Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli. Nucl Ac Res* **23**(13):3188–3193

Munoz ET, Bogarad LD and MW Deem (2004) Microarray and EST databset estimates of mRNA expression levels differ: The protein length versus expression curve for *C. elegans*. *BMC Genomics* **5**:30

Nagylaki T (1983) Evolution of a finite population under gene conversion. *Proc Natl Acad Sci USA* **80**:6278–6281

Nelson CE, Hersh BM and SB Carroll (2004) The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol* **5**(4): R25

Nielsen R (2005) Molecular signatures of natural selection. *Ann Rev Genetics* **39**:197–218

Ometto L, Stephan W and D De Lorenzo (2005) Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* **169**:1521–1527.

Ometto L, De Lorenzo D and W Stephan (2006) Contrasting patterns of sequence divergence and base composition between *Drosophila* introns and intergenic regions. *Biol Letters* **2**:604–607

Pál C, Papp B and LD Hurst (2001) Highly expressed genes in yeast evolve slowly. *Genetics* **158**:927–931

Papaceit M and E Juan (1998) Fate of dot chromosome genes in *Drosophila willistoni* and *Scaptodrosophila lebanonensis* determined by in sity hybridization. *Chrom Res* **6**(1):49–54

Parsch J (2003) Selective constraints on intron evolution in Drosophila. *Genetics* **165**:1843–1851

Parsch J (2004) Functional analysis of *Drosophila melanogaster* gene regulatory sequences by transgene coplacement. *Genetics* **168**:559–561

Patterson JT and WS Stone (1952) *Evolution in the Genus Drosophila.* MacMillan, New York

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes and SP Fodor (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* **91**(11):5022–5026

Petrov DA (2002) DNA loss and evolution of genome size in *Drosophila*. *Genetica*

115:81–91

Petrov DA, Lozovskaya ER and DL Hartl (1996) High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**:346–349

Petrov DA and DL Hartl (1997) Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila*. *Gene* **205**:279–289

Petrov DA and DL Hartl (1998) High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol Biol Evol* **15**(3):293–302

Petrov DA and DL Hartl (2000) Pseudogene evolution and natural selection for a compact genome. *Am Genet Assoc* **9**1:221–227

Plotkin JB, Dushoff J, Desai MM and HB Fraser (2006) Codon usage and selection on proteins. *J Mol Evol* **63**:635–653

Post LE and M Nomura (1980) DNA sequences from the *str* operon of *Escherichia coli*. *J Biol Chem* **255**:4660–4666

Powell JR (1997) *Progress and Prospects in Evolutionary Biology: The Drosophila Model* (Oxford Univ. Press, New York)

Powell JR, Moriyama EN (1997) Evolution of codon usage bias in *Drosophila*. *Proc Natl Acad Sci USA* **94**:7784–7790

Presgraves DC (2006) Intron length evolution in *Drosophila*. *Mol Biol Evol* **23**(11):2203–2213.

Ptak SE and DA Petrov (2002) How intron splicing affects the deletion and insertion profile in *Drosophila melanogaster*. *Genetics* **162**:1233–1244

Reenan RA (2005) Molecular determinants and guided evolution of species-specific RNA editing. *Nature* **434**:409–413

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, Couronne O, Hua S, Smith MA, Zhang P, Liu J, Bussemaker, HJ, van Batenburg MF, Howells SL, Scherer SE, Sodergren E, Matthews BB, Crosby MA, Schroeder AJ, Ortiz-Barrientos D, Rives CM, Metzker ML, Muzny DM, Scott G, Steffen D, Wheeler DA, Worley KC, Havlak P, Durbin KJ, Egan A, Gill R, Hume J, Morgan MB, Miner G, Hamilton C, Huang Y, Waldron L, Verduzco D, Clerc-Blankenburg KP, Dubchak I, Noor MA, Anderson

W, White KP, Clark AG, Schaeffer SW, Gelbart W, Weinstock GM and RA Gibbs (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res* **15**:1–18

Riddle NC, Shaffer CD and SCR Elgin (2009) A lot about a little dot – lessons learned from *Drosophila melanogaster* chromosome 4. *Biochem Cell Biol* **87**:229–241

Riddle NC and SCR Elgin (2006) The dot chromosome of *Drosophila*: Insights into chromatin states and their change over evolutionary time. *Chromosome Res* **14**:405–416

Rogic S, Montpetit B, Hoos HH, Mackworth AK, Ouellette BFF and P Hieter (2008) Correlation between the secondary structure of pre-mRNA introns and the efficiency of splicing in *Saccharomyces cerevisiae*. *BMC Genomics* **9**:355

Royce TE, Rozowsky JS and MB Gerstein (2007) Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics* **23**(8):988–997

Rozas J, Sanchez-DelBarrio JC and X Messeguer (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**(18):2496–2497

Schaeffer SW (2002) Molecular population genetics of sequence length diversity in the *Adh* region of *Drosophila pseudoobscura*. *Genet Res Camb* **80**:163–175

Schaeffer SW, Walthour CS, Toleno DM, Olek AT and EL Miller (2001) Protein variation in *Adh* and *Adh*-related in *Drosophila pseudoobscura*: Linkage disequilibrium between single nucleotide polymorphisms and protein alleles. *Genetics* **159**:673–687

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguadé M, Anderson WW, et al. (38 authors) (2008) Polytene chromosomal maps of 11 Drosophila species: The order of genomic scaffolds inferred from genetic and physical maps. *Genetics* **179**:1601–1655

Shabalina SA and NA Spiridonov (2004) The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology* **5**:105

Sharp PM and WH Li (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* **24**(1-2):28–38

Sharp PM and WH Li (1987) The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucl Ac Res* **15**:1281–1295

Sharp PM, Averof M, Lloyd AT, Matassi G and JF Peden (1995) DNA sequence evolution: the sounds of silence. *Phil Trans: Biol Sci* **349**(1329):241–247

Shendure J and H Li (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**(10):1135–1145

Shields DC, Sharp PM, Higgins DG and F Wright (1988) "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. *Mol Biol Evol* **5**(6):704–716

Singh ND, Davis JC and DA Petrov (2005a) Codon bias and noncoding GC content correlate negatively with recombination rate on the Drosophila X chromosome. *J Mol Evol* **61**:315–324

Singh ND, Davis JC and DA Petrov (2005b) X-linked genes evolve higher codon bias in *Drosophila* and *Caenorhabditis*. *Genetics* **171**:145–155

Singh ND, VL Bauer DuMont, MJ Hubisz, R Nielsen and CF Aquadro (2007) Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol Biol Evol* **24**(12):2687–2697

Singh ND, AM Larracuente and AG Clark (2008) Contrasting the efficacy of selection on the X and autosomes in Drosophila. *Mol Biol Evol* **25**(2):454–467

Sironi M, Menozzi G, Comi GP, Cagliani R, Bresolin N and U Pozzoli (2005) Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum Mol Genet* **14**(17):2533–2546

Slatkin M and RR Hudson (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**:555–562

Smith MW (1988) Structure of vertebrate genes: a statistical analysis implicating selection. *J Mol Evol* **27**:45–55

Sokal RR and FJ Rohlf (1995) *Biometry: The Principles and Practice of Statistics in Biological Research* (W.H. Freeman, New York)

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Harvard FlyBase curators, Berkeley Drosophila Genome Project, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SW, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM and M Kellis (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**(7167):219–232

Stephan W, Rodriguez VS, Zhou B and J Parsch (1994) Molecular evolution of the Metallothionein gene *Mtn* in the *melanogaster* species group: Results from *Drosophila ananassae*. *Genetics* **138**:135–143

Stone EA, Cooper GM and S Arend (2005) Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Ann Rev Genomics and Human Genetics* **6**:143–164

Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* **168**:373–381

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595

Thornton K (2003) libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**(17):2325–2327

Throckmorton LH (1982) The *virilis* species group. In: Ashburner M, Carson HL and JN Thompson (eds) *The Genetics and Biology of Drosophila* pp. 227–296. Academic Press, New York.

Vicario S, Moriyama EN and JR Powell (2007) Codon usage in twelve species of Drosophila. *BMC Evol Biol* **7**:226

Vieira J, McAllister BF and B Charlesworth (2001) Evidence for selection at the *fused1* locus of *Drosophila americana*. *Genetics* **158**:279–290

Vieira CP, Almeida A, Dias JD and J Vieira (2006) On the location of the gene(s) harbouring the advantageous variant that maintains the *X/4* fusion of *Drosophila americana*. *Genet Res Camb* **87**:163–174

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM and LA Pennacchio (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**(2):158–160.

Wang J, Keightley PD and T Johnson (2006) MCALIGN2: Faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. *BMC Bioinformatics* **7**:292

Warnecke T, Batada NN and LD Hurst (2008) The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genetics* **4**(11):e1000250

Warters M (1944) Chromosomal aberrations in wild populations of Drosophila. Univ. Texas Pub. 4445:129–174

Weber CC and LD Hurst (2010) Protein rates of evolution are predicted by double-strand break events, independent of crossing-over rates. *Genome Biol Evol* **2009**:340–349

Wright SI, Lauga B and D Charlesworth (2002) Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Mol Biol Evol* **19**:1407-1420

Wright SI and B Charlesworth (2004) The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168**:1071–1076

Wyckoff GJ, Malcom CM, Vallender EJ and BT Lahn (2005) A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends Genet* **21**(7):381–385

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**(5):555–556

Yang Z (2007) PAML4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**(8):1586–1591

Yi S and B Charlesworth (2000) Contrasting patterns of molecular evolution of the genes on the new and old sex chromosomes of *Drosophila miranda*. *Mol Biol Evol* **17**(5):703–717

Zeng K and B Charlesworth (2009) Estimating selection intensity on synonymous codon usage in a nonequilibrium population. *Genetics* **183**:651–662

Zeng K and B Charlesworth (2010) Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. *J Mol Evol* **70**:116–128

Zhang Y, Sturgill D, Parisi M, Kumar S and B Oliver (2007) Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* **450**(7167):233–238

Appendix

Appendix 2.1: Perl script used to extract introns from alignments of *D. miranda* BAC sequences with the *D. pseudoobscura* sequence in Chapter 2.

```
#!/usr/bin/perl -w
use strict;

###############################################################################
# read a fasta file, including gaps, into an array
# each element in the array is a hash with two keys:  "name" and "seq"
# "name" is the sequence name
# "seq" is an array where each element is a base in the sequence
# usage:  e.g. read_fasta_array($file, \@seqs);
###############################################################################
sub read_fasta_array
{
  my ($file, $seq) = @_;
  @$seq = ();               # wipe array contents

  open (INF, "$file") || die "read_fasta: cannot open $file for read, reason $!\n";
  my $name;
  my $seqno;
  while(defined(my $iline = <INF>)) {
    next if(($iline =~ /^\s*$/)||($iline =~ /^\s*#/));        # ignore blank and comment lines
    chomp $iline;                                  # remove trailing \n
    if ($iline =~ s/^>//) {                                  # check for angle bracket and remove if
present
      if(defined($seqno)) { $seqno++ }
      else { $seqno = 0; }
      $$seq[$seqno]{'name'} = $iline;                  # set sequence name
      next;
    }
    unless(defined($$seq[$seqno]{'name'})) { return(1); }  # text found before sequence name
    $iline =~ tr/A-Z/a-z/;                                   # lowercase sequence
    $iline =~ s/\s+//g;                            # remove white spaces from sequence line
    push(@{$$seq[$seqno]{'seq'}}, split(//, $iline));      # append array to sequence
  }
  close(INF);
  if(@$seq == 0) {
      warn "read_fasta_array: no sequences read\n";
      return(1);
  }
  return(0);
}

###############################################################################
```

```
# read_file_into_array -- reads a tab seperated file with header into
# an array where indexes in the array refer to lines in the file.
# each index in the array is a hash, where the hash keys are the names of the
# columns and values are the values in the columns for the appropriate line.
#
# this routine takes the file name and returns the array
################################################################################
sub read_file_into_array
{
	my ($file) = $_[0];

	open(INF, $file) || die "cannot open $file for read\n";
	chomp(my $line = <INF>);                                # read header
	my @header = split(/\t/, $line);

	my $i=0;
	my @array;
	while(my $line = <INF>) {
		chomp($line);
		next if($line =~ /^\s*$/);                  # skip blank lines
		my @tmp = split(/\t/, $line);
		foreach my $col (@header) {
			$array[$i]{$col} = shift(@tmp);
		}
		$i++;
	}
	close(INF);
	return(@array);
}

################################################################################
# extract_from_alignment -- extracts sequence from an alignment according
# to co-ordinates referring to the first sequence
################################################################################
sub extract_from_alignment
{
	my ($seqs, $start, $end) = @_;
	my @extract;

#	print "$start\t$end\n";

	# run through first sequence
	my $counter = 0;

	for(my $i=0; $i< @{$$seqs[0]{'seq'}} ; $i++) {

		$counter++ if($$seqs[0]{'seq'}[$i] =~ /[A-Za-z]/);
		if(($counter > $start)&&($counter <= $end)) {
#			print "$counter\t$$seqs[0]{'seq'}[$i]\t$$seqs[1]{'seq'}[$i]\n";
			for(my $j=0; $j<@$seqs; $j++) {
				$extract[$j] .= $$seqs[$j]{'seq'}[$i];
```

```
                }
            }
            last if($counter == $end);
        }
#       print "$extract[0]\n";
#       print "$extract[1]\n";
        return(@extract);
}

##############################################################################
# revcomp -- reverse complement of a sequence
##############################################################################
sub revcomp
{
        $_[0] =~ tr/ACGTRYacgtry/TGCAYRtgcayr/;
        $_[0] =  reverse($_[0]);
        return($_[0]);
}

##############################################################################
#       print_single_fasta  prints a single sequence in fasta format,
#  receives file name, sequence name and sequence
#       e.g. print_single_fasta($file, $name, $seqs{$name});
##############################################################################
sub print_single_fasta
{
  my ($file, $name, $seq) = @_;

        unless((defined($file))&&(defined($name))&&(defined($seq))) { return(1); }
        print $file ">$name\n";
        for(my $i=0; $i<length($seq); $i += 60) { print $file substr($seq, $i, 60), "\n"; }
        print $file "\n";
        return(0);
}

##############################################################################
#       get_scaler      -- when passed a phrase it prints phase until a valid scaler is entered.
#                               returns the scaler only if the input has a length greater than 1
#                       -- optionally pass a default value, which will be used if nothing is entered.
#       e.g. $string = get_scaler("enter a string");
#       or   $string = get_scaler("enter a string", "files");
##############################################################################
sub get_scaler
{
        for(my $i=0; $i<3; $i++) {
                if($i >= 1)                             { print STDERR "Error, "; }
                if(defined($_[0])) {
                        if(defined($_[1]))      { print STDERR "$_[0] [$_[1]]: "; }
                        else                                            { print STDERR "$_[0]: "; }
                }
                else {
```

```
                if(defined($_[1]))    { print STDERR "[$_[1]]: "; }
        }

        chomp(my $input=<STDIN>);
        unless(defined($input)) { die "get_scaler: input is undefined\n"; }

        if(defined($_[1])&&(length($input)==0)) { return($_[1]); }

        $input =~ s/^\s+//;          # remove leading and trailing white space
        $input =~ s/\s+$//;
        if(length($input)>0) { return($input); }
    }
    die "get_scaler: get_scaler failed 3 times\n";
}

################################################################################
#       create_or_wipe_dir -- checks if a directory exists, if so, asks if it should be wiped
#  if not, directory is created (inc parent directories if necessary)
#       e.g. check_dir($dir);
################################################################################
sub create_or_wipe_dir
{
    my $dir = $_[0];
    my $res = get_scaler("If directory \"$dir\" exists, wipe contents?", 'n');
    if(-d "$dir") {
        return(0) unless($res =~ /^y/i);
        system "rm -rf $dir";
    }
    die "create_or_wipe_dir: $dir exists but is not a directory\n" if(-e $dir);
    system "mkdir -p $dir";
    return(0);
}

################################################################################
# MAIN PROGRAM
################################################################################

# get file names from @ARGV
my $usage = "usage: extract-coding.pl <coordinate file> <alignment file> <results directory>";
my $coordfile = shift or die "$usage\n";
my $seqfile = shift or die "$usage\n";
my $outdir = shift or die "$usage\n";
create_or_wipe_dir($outdir);

# read fasta file
my @seqs;
read_fasta_array($seqfile, \@seqs);

# parse 1st sequence name
my ($tmp, $seqfile_start, $seqfile_end) = split(/_/, $seqs[0]{'name'});
```

```
# parse coordinate file
my @coords = read_file_into_array($coordfile);

# for each line from the coordinate file
for(my $i=0; $i<@coords; $i++) {
	print "$coords[$i]{'#name'}\n";

	# check whether alignment and reference strand are in the same direction
	my $strand = $coords[$i]{'strand'};
	my $frag_strand;
	if ($seqfile =~ /revcomp/) {
		$frag_strand = "-";
	}
	else {
		$frag_strand = "+";
	}
	my $samedir;
	if ($frag_strand eq $strand) {
		$samedir = "yes";
	} else {
		$samedir = "no";
	}
	print "$samedir\n";

	# arrays of intron start and end coordinates
	my @starts = split(/,/, $coords[$i]{'exonStarts'});
	my @ends   = split(/,/, $coords[$i]{'exonEnds'});

	# @introns is an array where each element represents an exon (which is an array
	# where each element is the sequence for each species
	# @int is an array where each element is the sequence for each species
	my @introns;
	my @int;

	# for each exon -- extract alignment and add to coding sequence
	for(my $j=0; $j<@starts; $j++) {
		# adjust numbers so they start at start of sequence
		$starts[$j] = $starts[$j] - $seqfile_start;
		$ends[$j] = $ends[$j] - $seqfile_start;
		print "extracting exon $j:  $starts[$j]\t$ends[$j]\n";

		@{$introns[$j]} = extract_from_alignment(\@seqs, $starts[$j], $ends[$j]);
		for(my $k=0; $k<@{$introns[$j]}; $k++) {
			$int[$k] .= $introns[$j][$k];
		}
	}

	# if opposite direction, reverse order of introns
	unless ($samedir =~ "yes") {
		print "Introns array reverted\n";
		@introns = reverse @introns;
```

```
	}

	open(OUT, ">$outdir/$coords[$i]{'#name'}") || die "cannot open outfile
$outdir/$coords[$i]{'#name'} for write\n";

	for(my $k=0; $k<@int; $k++) {
		unless ($samedir =~ "yes") {
		print "CDS reverse-complemented\n";
			$int[$k] = revcomp($int[$k]); }
		print "$seqs[$k]{'name'}\n";
		print_single_fasta(\*OUT, $seqs[$k]{'name'}, $int[$k]);
	}
	close(OUT);
}
```

Appendix 3.1: Results from the Mantel-Haenszel tests in Chapter 3.

Test	Nb of genes compared	Statistic MH χ^2	MH p-value	Common Odds Ratio
All/Synonymous	15	2.3003	0.1294	1.2885
First/Second	13	0.0038	0.9508	1.0281
First/Synonymous	13	0.6601	0.4165	1.1911
Non-First/Synonymous	15	0.0086	0.9260	1.0388
Short/Long	8	0.0242	0.8764	0.9115
Short/Synonymous	12	0.2396	0.6245	1.1899
Long/Synonymous	17	2.3739	0.1234	1.2697
Excluding singletons				
All/Synonymous	15	2.1963	0.1383	1.3446
First/Second	13	0.0050	0.9438	1.0156
First/Synonymous	13	0.1180	0.7312	1.1126
Non-First/Synonymous	15	1.7244	0.1891	1.3542
Short/Long	8	0.2863	0.5926	0.7578
Short/Synonymous	12	0.0273	0.8688	0.8884
Long/Synonymous	17	2.5049	0.1135	1.4004

Appendix 4.1: Table with details for all insertions and deletions for *D. americana* from Chapter 4, both complex and simple, with location on the intron alignment, indel size and frequency. Out of 296 indel events, 209 (70.6%) were complex indel events and 87 (29.4%) were simple indel events.

Gene	Intron	Start	End	Size	Polarised	Frequency	Category
cv	1	86	92	7	Insertion	0.071	Simple
cv	1	189	190	2	Deletion	0.071	Complex
cv	1	191	191	1	Deletion	0.143	Complex
cv	1	192	194	3	Insertion	0.929	Complex
cv	1	195	195	1	Insertion	0.929	Simple
cv	1	199	200	2	Deletion	0.071	Simple
l1Bi	1	51	61	11	Deletion	0.857	Simple
l1Bi	1	209	217	9	Deletion	0.214	Simple
l1Bi	1	239	248	10	Deletion	0.500	Complex
l1Bi	1	249	266	18	Deletion	0.786	Complex
l1Bi	1	267	272	6	Deletion	0.500	Complex
l1Bi	1	415	416	2	Insertion	0.786	Simple
l1Bi	1	523	526	4	Insertion	0.500	Simple
l1Bi	1	550	551	2	Insertion	0.929	Complex
l1Bi	1	552	561	10	Insertion	0.786	Complex
l1Bi	1	562	568	7	Insertion	0.429	Complex
l1Bi	1	569	594	26	Insertion	0.643	Complex
l1Bi	1	601	601	1	Insertion	0.071	Complex
l1Bi	1	602	607	6	Deletion	0.071	Complex
l1Bi	1	608	609	2	Deletion	0.143	Complex
l1Bi	1	610	653	44	Deletion	0.071	Complex
l1Bi	1	654	654	1	Deletion	0.500	Complex
l1Bi	1	655	656	2	Deletion	0.071	Complex
l1Bi	1	657	660	4	Insertion	0.429	Complex
l1Bi	1	661	672	12	Insertion	0.929	Complex
l1Bi	1	673	616	4	Insertion	0.857	Complex
l1Bi	1	677	712	36	Insertion	0.929	Complex
l1Bi	1	713	718	6	Deletion	0.071	Complex
l1Bi	1	719	719	1	Deletion	0.143	Complex
l1Bi	1	720	721	2	Deletion	0.071	Complex
l1Bi	1	722	722	1	Deletion	0.143	Complex
l1Bi	1	723	723	1	Deletion	0.071	Complex
l1Bi	1	724	724	1	Insertion	0.214	Complex
l1Bi	1	725	726	2	Deletion	0.143	Complex
l1Bi	1	727	728	2	Deletion	0.143	Complex
l1Bi	1	759	770	12	Insertion	0.214	Complex
l1Bi	1	851	851	1	Deletion	0.214	Simple
17292	1	82	82	1	Insertion	0.143	Simple
17292	1	113	118	6	Insertion	0.214	Complex
17292	1	119	120	2	Insertion	0.143	Complex
17292	1	121	122	2	Insertion	0.286	Complex

Gene	Intron	Start	End	Size	Polarised	Frequency	Category
17292	1	123	124	2	Insertion	0.429	Complex
17292	1	125	128	4	Deletion	0.143	Complex
17292	1	416	416	1	Deletion	0.929	Simple
17292	1	450	450	1	Deletion	0.071	Simple
17292	1	498	507	10	Deletion	0.857	Complex
17292	1	508	509	2	Insertion	0.071	Complex
17292	1	510	515	6	Deletion	0.929	Complex
17292	1	516	519	4	Deletion	0.857	Complex
17292	1	782	782	1	Deletion	0.143	Simple
17292	1	813	814	2	Deletion	0.143	Simple
17292	1	830	831	2	Insertion	0.071	Complex
17292	1	832	832	1	Insertion	0.143	Complex
17292	1	833	833	1	Insertion	0.500	Complex
17292	1	834	834	1	Deletion	0.286	Complex
17292	1	835	835	1	Deletion	0.143	Complex
17292	1	836	836	1	Deletion	0.071	Complex
17292	1	858	858	1	Deletion	0.929	Complex
17292	1	859	866	8	Insertion	0.071	Complex
Cyp28c1	1	121	121	1	Deletion	0.071	Complex
Cyp28c1	1	140	140	1	Insertion	0.071	Simple
Cyp28c1	1	142	143	2	Insertion	0.071	Simple
Cyp28c1	1	160	164	5	Deletion	0.286	Simple
Cyp28c1	1	310	310	1	Deletion	0.429	Simple
Cyp28c1	1	509	510	2	Insertion	0.071	Complex
Cyp28c1	1	511	511	1	Insertion	0.857	Complex
Cyp28c1	1	512	516	5	Insertion	0.929	Complex
Cyp28c1	1	521	521	1	Deletion	0.071	Complex
Cyp28c1	1	522	522	1	Insertion	0.929	Complex
Cyp28c1	1	524	524	1	Insertion	0.857	Complex
Cyp28c1	1	533	533	1	Insertion	0.929	Complex
Cyp28c1	1	534	534	1	Insertion	0.571	Complex
Cyp28c1	1	535	535	1	Insertion	0.500	Complex
Cyp28c1	1	536	536	1	Insertion	0.286	Complex
Cyp28c1	1	537	538	2	Insertion	0.143	Complex
Cyp28c1	1	539	539	1	Insertion	0.786	Complex
Cyp28c1	1	579	579	1	Deletion	0.286	Simple
Cyp28c1	2	28	28	1	Deletion	0.143	Simple
Cyp28c1	2	65	65	1	Insertion	0.786	Simple
Cyp28c1	2	117	122	6	Deletion	0.071	Simple
Cyp28c1	2	144	156	13	Deletion	0.286	Simple
Cyp28c1	2	167	173	7	Deletion	0.357	Complex
Cyp28c1	2	174	181	8	Deletion	0.071	Complex
dor	2	437	437	1	Insertion	0.714	Complex
eIF5	2	82	82	1	Deletion	0.538	Complex
eIF5	2	84	88	5	Deletion	0.538	Complex
eIF5	2	89	90	2	Insertion	0.462	Complex

Gene	Intron	Start	End	Size	Polarised	Frequency	Category
eIF5	2	91	94	4	Deletion	0.538	Complex
eIF5	2	96	104	9	Insertion	0.538	Simple
eIF5	2	442	446	5	Deletion	0.077	Simple
eIF5	2	476	486	11	Deletion	0.308	Simple
eIF5	2	501	503	3	Deletion	0.692	Complex
eIF5	2	504	510	7	Deletion	0.769	Complex
eIF5	2	511	514	4	Insertion	0.077	Complex
eIF5	2	515	518	4	Insertion	0.154	Complex
eIF5	2	519	535	17	Deletion	0.692	Complex
GJ16746	1	7	7	1	Deletion	0.846	Simple
GJ16746	1	22	22	1	Deletion	0.538	Simple
GJ16746	1	27	27	1	Deletion	0.077	Simple
GJ16746	1	31	31	1	Deletion	0.077	Simple
GJ16746	1	37	37	1	Deletion	0.077	Simple
GJ16746	1	50	50	1	Deletion	0.077	Simple
GJ16746	1	119	121	3	Deletion	0.615	Simple
GJ16746	1	239	245	7	Insertion	0.077	Complex
GJ16746	1	246	251	6	Deletion	0.923	Complex
GJ16746	1	260	272	13	Deletion	0.308	Simple
GJ16746	1	354	365	12	Deletion	0.786	Simple
GJ16746	1	418	418	1	Insertion	0.071	Simple
GJ16746	2	34	34	1	Insertion	0.071	Simple
rad	1	232	233	2	Insertion	0.143	Complex
rad	2	121	123	3	Insertion	0.692	Complex
rad	2	124	129	6	Insertion	0.615	Complex
rad	2	130	139	10	Insertion	0.077	Complex
rad	2	148	154	7	Insertion	0.692	Complex
rad	2	155	162	8	Insertion	0.615	Complex
rad	2	163	167	5	Deletion	0.308	Complex
rad	2	168	169	2	Insertion	0.692	Complex
rad	2	197	206	10	Insertion	0.615	Simple
rad	2	248	271	24	Insertion	0.923	Complex
rad	2	272	280	9	Insertion	0.769	Complex
rad	2	281	307	27	Insertion	0.923	Complex
rad	2	323	334	12	Insertion	0.385	Simple
rad	2	336	336	1	Insertion	0.923	Complex
rad	2	337	337	1	Insertion	0.615	Complex
rad	2	338	338	1	Insertion	0.692	Complex
rad	2	457	472	16	Deletion	0.538	Simple
rad	2	525	529	5	Deletion	0.077	Simple
csw	1	12	18	7	Deletion	0.071	Simple
csw	1	52	61	10	Deletion	0.571	Simple
csw	2	67	67	1	Deletion	0.143	Simple
sog	4	43	51	9	Insertion	0.786	Simple
sog	4	79	80	2	Insertion	0.071	Complex
sog	4	81	82	2	Insertion	0.286	Complex

Gene	Intron	Start	End	Size	Polarised	Frequency	Category
sog	4	131	133	3	Deletion	0.214	Simple
sog	4	180	207	28	Deletion	0.571	Simple
sog	4	231	232	2	Insertion	0.071	Complex
sog	4	233	234	2	Insertion	0.214	Complex
sog	4	374	376	3	Deletion	0.071	Complex
sog	4	377	381	5	Insertion	0.929	Complex
sog	4	382	384	2	Insertion	0.643	Complex
sog	4	385	385	1	Insertion	0.286	Complex
sog	4	386	386	1	Deletion	0.714	Complex
sog	4	376	386	11	Deletion	0.286	Complex
sog	4	401	404	4	Deletion	0.286	Simple
sog	4	418	427	10	Deletion	0.143	Simple
sog	4	591	591	1	Insertion	0.929	Simple
singed	6	100	104	5	Deletion	0.643	Complex
singed	6	105	129	25	Deletion	0.643	Complex
singed	6	130	132	3	Insertion	0.214	Complex
singed	6	133	225	93	Insertion	0.357	Complex
singed	6	237	250	14	Insertion	0.929	Complex
singed	6	251	255	5	Insertion	0.786	Complex
singed	6	256	274	19	Insertion	0.929	Complex
singed	6	364	366	3	Deletion	0.071	Simple
singed	6	409	411	3	Deletion	0.214	Simple
singed	6	440	444	5	Deletion	0.286	Complex
singed	6	445	449	5	Insertion	0.071	Complex
singed	6	450	460	11	Deletion	0.286	Complex
singed	6	508	508	1	Deletion	0.071	Simple
singed	6	584	610	27	Deletion	0.357	Simple
singed	6	624	630	7	Deletion	0.071	Complex
singed	6	631	633	3	Insertion	0.929	Complex
singed	6	640	647	8	Deletion	0.071	Complex
singed	6	724	727	4	Insertion	0.929	Simple
shibire	6	31	32	2	Deletion	0.286	Complex
shibire	6	33	34	2	Deletion	0.071	Complex
shibire	6	55	55	1	Deletion	0.214	Complex
shibire	6	56	58	3	Deletion	0.143	Complex
shibire	6	59	60	2	Deletion	0.857	Complex
shibire	6	95	96	2	Insertion	0.071	Complex
shibire	6	97	101	5	Deletion	0.857	Complex
shibire	6	108	113	6	Deletion	0.857	Complex
shibire	6	114	115	2	Insertion	0.143	Complex
shibire	6	116	116	1	Deletion	0.857	Complex
shibire	6	183	184	2	Insertion	0.286	Complex
shibire	6	185	186	2	Deletion	0.143	Complex
shibire	6	202	203	2	Insertion	0.214	Complex
shibire	6	276	277	2	Deletion	0.071	Complex
shibire	6	278	279	2	Deletion	0.714	Complex

Gene	Intron	Start	End	Size	Polarised	Frequency	Category
shibire	6	280	283	4	Deletion	0.857	Complex
shibire	6	284	285	2	Deletion	0.929	Complex
shibire	6	305	327	23	Insertion	0.929	Simple
shibire	6	337	352	16	Insertion	0.929	Simple
shibire	6	369	372	4	Insertion	0.143	Simple
shibire	6	492	497	6	Insertion	0.929	Complex
shibire	6	498	500	3	Deletion	0.071	Complex
shibire	6	595	596	2	Insertion	0.857	Simple
shibire	6	634	634	1	Insertion	0.214	Simple
shibire	6	730	731	2	Insertion	0.357	Complex
shibire	6	732	732	1	Insertion	0.500	Complex
shibire	6	733	733	1	Insertion	0.857	Complex
shibire	6	756	764	9	Deletion	0.846	Complex
shibire	6	723	724	4	Deletion	0.923	Complex
shibire	6	728	731	2	Deletion	0.077	Complex
shibire	6	836	836	1	Insertion	0.538	Complex
phl	4	90	134	45	Deletion	0.143	Complex
phl	4	135	138	4	Deletion	0.286	Complex
phl	4	139	201	63	Deletion	0.143	Complex
phl	4	203	206	4	Deletion	0.143	Complex
phl	4	207	210	4	Deletion	0.643	Complex
phl	4	211	230	20	Deletion	0.143	Complex
phl	4	231	233	3	Deletion	0.214	Complex
phl	4	234	244	11	Insertion	0.071	Complex
phl	4	245	255	11	Deletion	0.214	Complex
phl	4	256	259	4	Deletion	0.143	Complex
phl	4	260	291	32	Insertion	0.857	Complex
phl	4	292	292	1	Insertion	0.571	Complex
phl	4	293	293	1	Insertion	0.071	Complex
phl	4	294	354	61	Insertion	0.857	Complex
phl	4	392	397	6	Insertion	0.929	Complex
phl	4	398	449	52	Insertion	0.214	Complex
phl	4	450	450	1	Insertion	0.143	Complex
phl	4	451	531	81	Insertion	0.214	Complex
phl	4	532	541	10	Insertion	0.929	Complex
phl	4	575	575	1	Insertion	0.214	Simple
phl	4	591	640	50	Insertion	0.071	Simple
phl	4	652	657	6	Deletion	0.143	Simple
phl	4	662	709	48	Deletion	0.071	Complex
phl	4	856	856	1	Insertion	0.286	Simple
phl	4	864	864	1	Insertion	0.500	Complex
phl	4	865	865	1	Insertion	0.857	Complex
phl	4	906	906	1	Insertion	0.429	Simple
phl	4	938	939	2	Insertion	0.214	Simple
phl	4	944	946	3	Insertion	0.714	Complex
phl	4	947	951	5	Insertion	0.786	Complex

Gene	Intron	Start	End	Size	Polarised	Frequency	Category
phl	4	952	966	15	Insertion	0.214	Complex
phl	4	989	1010	22	Insertion	0.786	Complex
phl	4	1011	1012	2	Deletion	0.214	Complex
phl	4	1013	1019	7	Deletion	0.786	Complex
phl	4	1020	1024	5	Deletion	0.500	Complex
phl	4	1025	1027	3	Insertion	0.500	Complex
phl	4	1192	1194	3	Deletion	0.071	Simple
mew	5	72	76	5	Deletion	0.071	Simple
mew	5	185	192	8	Deletion	0.071	Complex
mew	5	193	207	15	Deletion	0.571	Complex
mew	5	208	236	29	Deletion	0.071	Complex
mew	5	237	240	4	Insertion	0.929	Complex
mew	5	241	247	7	Deletion	0.071	Complex
mew	5	248	248	1	Deletion	0.286	Complex
mew	5	281	324	44	Deletion	0.429	Complex
mew	5	325	330	6	Insertion	0.571	Complex
mew	5	331	340	10	Insertion	0.429	Complex
mew	5	341	421	81	Insertion	0.571	Complex
mew	5	422	422	1	Insertion	0.500	Complex
mew	5	423	479	57	Insertion	0.571	Complex
mew	5	480	480	1	Insertion	0.429	Complex
mew	5	481	512	32	Insertion	0.571	Complex
mew	5	549	561	13	Insertion	0.714	Simple
mew	5	562	565	4	Deletion	0.286	Simple
mew	5	661	670	10	Deletion	0.071	Simple
mew	5	687	687	1	Insertion	0.143	Simple
mew	5	705	708	4	Deletion	0.071	Simple
mew	5	846	846	1	Insertion	0.857	Simple
mew	5	877	882	6	Insertion	0.071	Simple
pros28.1	1	46	46	1	Deletion	0.214	Simple
Yp1	2	18	18	1	Insertion	0.071	Simple
hep	1	84	86	3	Deletion	0.429	Complex
hep	1	87	87	1	Deletion	0.571	Complex
hep	1	88	90	3	Deletion	0.643	Complex
hep	1	91	92	2	Deletion	0.714	Complex
hep	1	93	94	2	Deletion	0.929	Complex
hep	1	95	104	10	Insertion	0.500	Simple
hep	1	105	109	5	Deletion	0.500	Simple
hep	1	208	217	10	Insertion	0.071	Complex
hep	1	218	219	2	Insertion	0.214	Complex
hep	1	220	220	1	Insertion	0.429	Complex
hep	1	221	221	1	Deletion	0.571	Complex
hep	1	222	223	2	Deletion	0.286	Complex
hep	1	224	225	2	Deletion	0.071	Complex
hep	1	266	266	1	Deletion	0.286	Simple
hep	1	436	445	10	Deletion	0.071	Simple

Gene	Intron	Start	End	Size	Polarised	Frequency	Category
hep	1	458	461	4	Deletion	0.214	Complex
hep	1	462	463	2	Insertion	0.571	Complex
hep	1	464	466	3	Insertion	0.214	Complex
hep	1	467	468	2	Insertion	0.214	Complex
eIF5	1	78	79	2	Insertion	0.917	Simple
sol	1	76	79	4	Deletion	0.923	Complex
sol	1	80	83	4	Deletion	0.769	Complex
sol	1	84	85	2	Deletion	0.692	Complex
sol	1	86	87	2	Deletion	0.462	Complex
sol	1	88	89	2	Deletion	0.308	Complex
sol	1	90	91	2	Deletion	0.077	Complex
sol	1	212	213	2	Insertion	0.077	Complex
sol	1	214	217	4	Insertion	0.923	Complex
sol	1	219	219	1	Insertion	0.846	Complex
sol	1	222	222	1	Deletion	0.154	Complex
sol	1	354	354	1	Insertion	0.154	Simple
sol	1	356	357	2	Deletion	0.077	Simple
sol	1	360	360	1	Deletion	0.077	Simple
sol	1	480	480	1	Deletion	0.154	Simple
sol	1	481	481	1	Insertion	0.385	Simple
sol	1	547	547	1	Deletion	0.077	Complex
sol	1	548	549	2	Deletion	0.154	Complex
sol	1	550	550	1	Deletion	0.615	Complex
sol	1	551	551	1	Deletion	0.769	Complex
sol	1	643	643	1	Insertion	0.231	Simple

Appendix 5.1: Table of codon usage used for *D. americana* in Chapter 5, from the table for *D. virilis* obtained by Betancourt et al. (2009). P indicates the preferred codons, there are 21 preferred codons in this table.

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe P	UCC Ser P	UAC Tyr P	UGC Cys P
UUA Leu	UCA Ser	UAA Stop *	UGA Stop *
UUG Leu P	UCG Ser P	UAG Stop *	UGG Trp -
CUU Leu	CCU Pro	CAU His	CGU Arg P
CUC Leu	CCC Pro P	CAC His P	CGC Arg P
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu P	CCG Pro	CAG Gln P	CGG Arg
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile P	ACC Thr P	AAC Asn P	AGC Ser
AUA Ile	ACA Thr	AAA Lys	AGA Arg
AUG Met -	ACG Thr	AAG Lys P	AGG Arg
GUU Val	GCU Ala	GAU Asp P	GGU Gly
GUC Val	GCC Ala P	GAC Asp	GGC Gly P
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG Val P	GCG Ala	GAG Glu P	GGG Gly